La struttura logica della causalità controfattuale

Prolegomeni: la questione causale e il giro controfattuale

Vi è, nella storia della filosofia occidentale, un interrogativo che resiste con ostinazione inusitata a qualsiasi tentativo di liquidazione definitiva: che cosa significa dire che un evento ne ha *causato* un altro? La domanda non è oziosa né meramente speculativa. Essa innerva la pratica scientifica, orienta il ragionamento giuridico, struttura la deliberazione morale e, non da ultimo, costituisce il nucleo epistemico attorno a cui si organizza la comprensione ordinaria del mondo fisico. Eppure, nonostante l'urgenza e la pervasività del concetto, una teoria pienamente soddisfacente della causalità tarda a farsi avanti con la grazia che ci si potrebbe attendere da un tema così frequentato.

La difficoltà non è contingente. Deriva, in larga misura, da una tensione strutturale che percorre tutta la tradizione: da un lato, si avverte l'esigenza di ancorare la causalità a qualcosa di empiricamente robusto, di verificabile, di inscritto nel tessuto osservabile degli accadimenti; dall'altro, si constata che nessuna successione regolare di eventi, per quanto sistematica e replicabile, riesce a catturare integralmente il senso di quella relazione di *produzione* o di *dipendenza* che il concetto di causa sembra reclamare come sua proprietà inalienabile. Hume ha nominato questa tensione con insuperabile chiarezza: ciò che l'esperienza effettivamente offre è la contiguità spaziotemporale e la successione costante; ciò che la mente vi sovrappone — la connessione necessaria — è una proiezione psicologica, non un dato del mondo esterno. La via regolaristica si è dunque mostrata, sin dall'inizio, insufficiente a render conto della modalità causale, a meno di non rinunciare alla modalità tout court, scelta che la maggior parte dei filosofi ha trovato inaccettabile.

Il giro decisivo verso l'analisi controfattuale risponde esattamente a questa insoddisfazione. L'idea di fondo è di una semplicità quasi elementare: un evento c causa un evento e se e soltanto se, qualora c non fosse accaduto, e non sarebbe accaduto. Il linguaggio del condizionale controfattuale — "se non fosse stato così, non sarebbe accaduto quest'altro" — emerge come il veicolo semantico privilegiato attraverso cui la relazione causale viene articolata. Non si tratta di una novità assoluta: lo stesso Hume, nel momento in cui offre la sua seconda definizione di causa, fa ricorso proprio a questa struttura ipotetica; e Mill, nella sua analisi dei metodi della ricerca scientifica, lascia trasparire la stessa intuizione. Ma è soltanto nel ventesimo secolo inoltrato, e segnatamente con David Lewis, che questa intuizione viene elevata

a teoria formale rigorosa, dotata di una semantica precisa e di un apparato tecnico all'altezza della complessità del problema.

La svolta lewisiana degli anni Settanta si inserisce in un contesto filosofico particolarmente fertile: la logica modale e la semantica dei mondi possibili, inaugurata da Saul Kripke e sviluppata da numerosi autori, offre gli strumenti formali per trattare enunciati ipotetici e controfattuali con la stessa precisione con cui la logica classica tratta gli enunciati assertori. Lewis coglie questa opportunità e costruisce, a partire da essa, una teoria della causalità che è al contempo una teoria del linguaggio controfattuale, una metafisica delle disposizioni e una topologia dello spazio dei mondi possibili.

Qualche decennio più tardi, il contributo di Judea Pearl inaugurerà un secondo capitolo altrettanto radicale. Pearl non proviene dalla tradizione della filosofia analitica, ma da quella dell'informatica, dell'intelligenza artificiale e della statistica: la sua formazione lo porta a costruire una teoria della causalità fondata non su mondi possibili e relazioni di somiglianza, bensì su modelli strutturali, grafi aciclici orientati e un calcolo formale degli interventi. Il risultato è una concezione in cui la causalità viene identificata con la struttura delle dipendenze funzionali tra variabili in un sistema, e i controfattuali vengono derivati, piuttosto che assunti come primitivi, da tale struttura mediante procedure di manipolazione simbolica dei modelli.

Tra i due approcci — quello lewisiano e quello pearliano — esiste una corrispondenza che è, al tempo stesso, illuminante e non banale. Entrambi si fondano sulla nozione di controfattuale; entrambi la trattano come lo strumento primario dell'analisi causale; entrambi aspirano a una formalizzazione rigorosa. Eppure le architetture teoriche divergono profondamente nei presupposti ontologici, nei primitivi formali, nelle applicazioni e nelle difficoltà residue. Comprendere questa convergenza e questa divergenza è l'obiettivo centrale della presente monografia.

Il percorso si snoda come segue. Nella Parte Seconda si espone la teoria lewisiana nella sua versione matura, con particolare attenzione all'apparato semantico dei mondi possibili, alla struttura dell'ordinamento di somiglianza e alle condizioni di verità dei condizionali controfattuali. Nella Parte Terza si affronta il quadro pearliano dei modelli causali strutturali, con le sue equazioni strutturali, i grafi aciclici orientati e il triplice gradino epistemico della scaletta causale. Nella Parte Quarta si analizza il calcolo degli interventi — il *do-calculus* — come strumento formale per il transito dal piano osservativo a quello interventistico e, da quest'ultimo, al piano propriamente controfattuale. La Parte Quinta è dedicata a un confronto critico tra le due architetture, alla ricognizione dei problemi irrisolti e alle prospettive di integrazione o superamento.

II. David Lewis e la semantica dei mondi possibili

II.1 — Il controfattuale come primitivo semantico

La ricerca di David Lewis sulla causalità controfattuale prende avvio da una domanda semantica: quali sono le condizioni di verità degli enunciati condizionali che le logiche standard — tanto quella classica quanto quella modale elementare — non riescono a catturare adeguatamente? I condizionali controfattuali, cioè quei condizionali la cui protasi descrive una situazione contraria al fatto, presentano un comportamento logico che li distingue nettamente dai condizionali materiali e da quelli di necessità stretta.

Si consideri un caso elementare. L'enunciato "se questo fiammifero fosse stato strofinato, si sarebbe acceso" esprime qualcosa di preciso sul comportamento del fiammifero nelle circostanze date: è bagnato, o difettoso, o in un ambiente privo di ossigeno. Eppure la protasi è falsa — il fiammifero non è stato strofinato — e il condizionale materiale, che è vero ogni volta che la protasi è falsa, non discrimina affatto tra questo enunciato sensato e l'assurdo "se questo fiammifero fosse stato strofinato, la luna sarebbe diventata di formaggio". Entrambi sarebbero egualmente veri in logica classica, il che mostra con evidenza che il condizionale materiale non coglie ciò che si vuole dire.

Analogamente, i condizionali di necessità stretta — del tipo "è necessario che se p allora q", ovvero $\Box(p \rightarrow q)$ nella notazione modale — sono troppo forti: implicano che la connessione valga in tutti i mondi possibili, laddove il controfattuale si riferisce a un sottoinsieme ristretto di mondi, quelli sufficientemente simili al mondo attuale. Lewis risolve questa difficoltà introducendo un terzo operatore condizionale, l'operatore di *would-conditional* $\Box\!\!\rightarrow$, dotato di una semantica propria irriducibile a quelle preesistenti.

II.2 — La struttura formale: mondi, somiglianza, verità

Sia $\mathbf{W}$ un insieme non vuoto di mondi possibili. Sia $\mathbf{w}$ il mondo attuale, elemento di $\mathbf{W}$. Per ogni mondo $w \in \mathbf{W}$, si definisce una relazione di preordine totale $\leq_w$ su $\mathbf{W}$, detta relazione di *somiglianza* o di *vicinanza* centrata in w, con la proprietà che $w \leq_w v$ per ogni $v \in \mathbf{W}$ (il mondo attuale è almeno altrettanto vicino a se stesso quanto qualsiasi altro mondo) e che w è l'unico elemento minimo di $\leq_w$ (il mondo attuale è più vicino a se stesso che a qualunque altro mondo distinto). Questa struttura — l'insieme $\mathbf{W}$ munito della famiglia di preordini $\{\leq_w : w \in \mathbf{W}\}$ — costituisce il *sistema di sfere* lewisiano, per riprendere la rappresentazione grafica che Lewis stesso adotta.

La valutazione semantica di una formula proposizionale φ in un mondo w è standard: si assume una funzione di valutazione $\mathbf{V} : \mathrm{Prop} \times \mathbf{W} \rightarrow \{0, 1\}$ che assegna a ogni formula atomica e a ogni mondo un valore di verità. Le connettive classiche vengono interpretate nel modo usuale. L'interesse si concentra sull'operatore condizionale $\Box\!\!\rightarrow$.

Le condizioni di verità di $\varphi \Box\!\!\rightarrow \psi$ al mondo w sono le seguenti:

$\varphi \Box\!\!\rightarrow \psi$ è vero in w se e soltanto se: (i) non esiste nessun mondo in $\mathbf{W}$ in cui φ sia vera (caso di antecedente impossibile, controfattuale vacuamente vero), oppure (ii) esiste un mondo u in $\mathbf{W}$ tale che φ è vera in u, e per ogni mondo v con $v \leq_w u$ in cui φ sia vera, ψ è pure vera in v.

Equivalentemente, nella versione più compatta che fa uso dell'insieme dei mondi-φ più vicini a w:

Sia $\mathbf{f}_\varphi(w) = \{u \in \mathbf{W} : \varphi \text{ è vera in } u \text{ e } \neg\exists v(v <_w u \wedge \varphi \text{ è vera in } v)\}$ l'insieme dei mondi φ-veri più vicini a w (il *selettore* associato all'antecedente φ nel mondo w). Allora:

$\varphi \mathrel{\square\!\!\rightarrow} \psi$ è vera in $w \Leftrightarrow$ per ogni $u \in \mathbf{f}_\varphi(w)$, ψ è vera in u.

Questa formulazione presuppone, per buona definizione, l'esistenza di un insieme non vuoto di mondi φ-veri più vicini, condizione che Lewis garantisce mediante un assioma di *Limite* (o, in alternativa, adottando l'assioma di *Strettezza dell'Unicità*, che afferma l'esistenza di un unico mondo più vicino per ogni antecedente non contraddittorio). La questione dell'assioma di Limite — se sia necessario, sufficiente, o se si possa fare a meno di esso accettando ordinamenti privi di minimo — è stata oggetto di discussione tecnica considerevole, che Lewis stesso non ha eluso.

La logica risultante, che Lewis denota **VC** nella classificazione che propone nella monografia *Counterfactuals* del 1973, è caratterizzata dai seguenti schemi di validità (tra i principali):

Identità: $\varphi \mathrel{\square\!\!\rightarrow} \varphi$ (ogni mondo φ-vero soddisfa φ)

Modus Ponens controfattuale: da $\varphi \mathrel{\square\!\!\rightarrow} \psi$ e φ, si deriva ψ

Strengthening dell'antecedente (non valido): da $\varphi \mathrel{\square\!\!\rightarrow} \psi$ non si può derivare $(\varphi \wedge \chi) \mathrel{\square\!\!\rightarrow} \psi$. Questo punto è cruciale: aggiungere informazioni all'antecedente può invalidare la conseguenza, il che rispecchia il comportamento intuitivo dei controfattuali. Se fosse vero "se accendessi il fuoco, mi riscalderei", non è detto che sia vero "se accendessi il fuoco mentre sono immerso in un lago ghiacciato, mi riscalderei".

Transitività (non valida in generale): da $\varphi \mathrel{\square\!\!\rightarrow} \psi$ e $\psi \mathrel{\square\!\!\rightarrow} \chi$ non si può derivare $\varphi \mathrel{\square\!\!\rightarrow} \chi$. Questo vale in logica classica ma fallisce nei controfattuali, come mostra il fatto che "se Bizet e Verdi fossero stati connazionali, Bizet sarebbe stato italiano" e "se Bizet fosse stato italiano, Bizet sarebbe stato del paese di Verdi" non implicano "se Bizet e Verdi fossero stati connazionali, Bizet sarebbe stato del paese di Verdi" (il quale non segue perché la connazionalità potrebbe aver reso Verdi francese anziché Bizet italiano).

Contraposizione (non valida): da $\varphi \mathrel{\square\!\!\rightarrow} \psi$ non si deriva $\neg\psi \mathrel{\square\!\!\rightarrow} \neg\varphi$, per ragioni analoghe alla mancanza di transitività.

Centering forte: $\varphi \wedge \psi$ implica $\varphi \mathrel{\square\!\!\rightarrow} \psi$ (se φ e ψ sono entrambi veri nel mondo attuale, allora nel mondo attuale stesso — il più vicino — φ implica ψ).

II.3 — La teoria della causalità per dipendenza controfattuale

Il passaggio dalla semantica dei controfattuali alla teoria della causalità avviene in Lewis attraverso il concetto di *dipendenza controfattuale*. Siano *c* ed *e* due eventi distinti — Lewis intende per evento la realizzazione di una proprietà da parte di un oggetto in un tempo e un luogo determinati, entità su cui non occorre qui soffermarsi nei dettagli della sua ontologia degli eventi. Si dicono che *e dipende controfattualmente da c* se e soltanto se il seguente biconditionale controfattuale è vero nel mondo attuale:

$$O(c) \mathrel{\square\!\!\rightarrow} O(e) \wedge \neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e)$$

dove O(x) denota il proposizione che l'evento x si verifica. In parole: e si verifica se e soltanto se c si verifica, dove il "se e soltanto se" è letto in senso controfattuale.

Nella pratica, poiché di norma sia *c* che *e* si verificano effettivamente (altrimenti non si parlerebbe di causalità), il primo congiunto è automaticamente vero per il principio di Centering forte. Il peso teorico ricade dunque interamente sul secondo: $\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e)$. Se c non si fosse verificato, e non si sarebbe verificato.

La *dipendenza causale* viene quindi identificata con la dipendenza controfattuale tra eventi effettivi: se c ed e si verificano entrambi, e se $\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e)$ è vero, allora c è *causalmente rilevante* per e in modo diretto. La *causalità* in senso pieno è poi definita come chiusura transitiva della dipendenza causale diretta: c causa e se e soltanto se esiste una catena di eventi $c = e_0, e_1, \ldots, e_n = e$ tale che ciascun e_{i+1} dipende controfattualmente da e_i.

Formalmente, chiamando **D** la relazione di dipendenza controfattuale diretta tra coppie di eventi effettivi, la relazione causale **C** è il suo involucro transitivo:

$$c \mathbf{C} e \Leftrightarrow \exists\, n \geq 1,\ \exists\, e_0, e_1, \ldots, e_n \in \mathbf{Events} : e_0 = c \wedge e_n = e \wedge \forall\, i \in \{0,\ldots,n-1\} : e_i \mathbf{D} e_{i+1}$$

Questa struttura a catena è introdotta da Lewis specificamente per fronteggiare i casi di *preemption* (o anticipazione causale), in cui due potenziali cause si presentano simultaneamente ma una "batte sul tempo" l'altra. Nel caso della preemption — che Lewis distingue dalla *overdetermination* (sovradeterminazione) — la causa effettiva non è tale che, togliendo solo essa, l'effetto non si produrrebbe (perché la causa di riserva interverrebbe), ma lo è in una catena di dipendenze più fine.

II.4 — L'ordinamento di somiglianza e la sua indeterminatezza

La teoria lewisiana ottiene la sua precisione tecnica a prezzo di un impegno metafisico non trascurabile: l'ordinamento di somiglianza tra mondi possibili deve soddisfare condizioni specifiche affinché l'analisi causale produca risultati plausibili. In particolare, la domanda "quali sono i mondi più vicini al mondo attuale in cui c non si verifica?" deve avere una risposta determinata e coerente.

Nel saggio del 1979 *Counterfactual Dependence and Time's Arrow*, Lewis elabora una gerarchia di priorità nell'ordinamento di somiglianza, pensata per dar conto del fatto che i controfattuali si "propagano" verso il futuro, non verso il passato (l'asimmetria

temporale della causalità). La gerarchia è la seguente, in ordine decrescente di importanza:

Prima priorità: evitare grandi "miracoli" — deviazioni dalla leggi di natura su scala ampia e pervasiva. Seconda priorità: massimizzare la regione di perfetta coincidenza con il mondo attuale (mantenere invariate quante più leggi di natura possibili in quante più regioni spaziotemporali possibile). Terza priorità: evitare piccoli "miracoli" — deviazioni localizzate ma comunque contrarie alle leggi. Quarta priorità: massimizzare la similarità approssimativa nei particolari.

Questa gerarchia ha importanti conseguenze. Consideriamo il classico esempio di Nixon e la bomba nucleare: se Nixon avesse premuto il pulsante, si sarebbe scatenata la guerra nucleare? Per rispondere occorre determinare quale sia il mondo più vicino al nostro in cui Nixon preme il pulsante. Lewis sosterrebbe che tale mondo richiede un piccolo miracolo nel passato prossimo a quell'azione (una deviazione localizzata dal decorso causale del mondo attuale), mantiene immutata tutta la storia precedente di quel passato prossimo, e si propaga poi causalmente verso un futuro di guerra nucleare. Il mondo alternativo in cui la storia successiva all'azione di Nixon restasse pacifica richiederebbe invece un secondo miracolo (la non-esplosione delle bombe), che è meno simile al nostro di quanto lo sia un mondo con la guerra.

Questa articolazione del concetto di somiglianza è stata criticata con argomenti acuti da Jonathan Bennett e da altri: l'ordinamento di Lewis sembra costruito ad hoc per giustificare le nostre intuizioni sui controfattuali, piuttosto che derivato da un principio indipendente di somiglianza tra mondi. In altri termini, si sospetta una certa circolarità: stabiliamo quali mondi sono più simili in funzione di ciò che vogliamo che i controfattuali dicano, anziché usare la somiglianza come base indipendente per determinare il valore di verità dei controfattuali. Lewis è consapevole di questa obiezione e ritiene che l'ordinamento di somiglianza sia effettivamente guidato da intuizioni substantive sulla struttura del mondo, ma la questione rimane aperta.

II.5 — Preemption, overdetermination e le revisioni del 2000

La teoria lewisiana originale — quella basata sulla semplice dipendenza controfattuale diretta o sulla catena di dipendenze — incontra serie difficoltà in presenza di casi strutturalmente complessi. I due scenari canonici sono la *preemption* e l'*overdetermination*.

Nell'overdetermination simmetrica, due cause indipendenti si verificano entrambe, ognuna da sola sufficiente a produrre l'effetto. Esempio: due cecchini sparano simultaneamente alla vittima; entrambi i proiettili colpiscono il bersaglio in modo indipendentemente letale. In questo caso, nessun controfattuale della forma ¬O(c_i) □→ ¬O(e) è vero, poiché togliendo il contributo di uno solo dei cecchini l'altro avrebbe comunque causato la morte. Eppure l'intuizione vorrebbe che entrambi i cecchini siano cause della morte. La teoria basata sulla dipendenza controfattuale fallisce qui in modo eclatante.

Nella preemption *tardiva* (late preemption), il problema è diverso: una causa effettiva c_1 e una causa potenziale c_2 si presentano, ma c_2 viene "disattivata" dal processo

causale innescato da c_1 prima di poter produrre l'effetto. Siccome c_2 non avrebbe avuto il tempo di intervenire qualora c_1 non si fosse verificata, il controfattuale $\neg O(c_1) \; \Box\!\!\rightarrow \neg O(e)$ è falso (perché c_2 sarebbe intervenuta), e neppure la catena di dipendenze riesce a identificare c_1 come causa.

Lewis ha risposto a queste difficoltà in modo evolutivo. Nella prima versione della teoria, i casi di preemption erano gestiti attraverso le catene di dipendenza. Nei decenni successivi, il problema si è mostrato refrattario a soluzioni completamente soddisfacenti. Il saggio *Causation as Influence* del 2000 rappresenta la rielaborazione più ambiziosa: Lewis vi introduce un concetto più raffinato di dipendenza causale, fondato non sulla semplice occorrenza o non-occorrenza dell'effetto, ma sulla sua *forma*, il suo *tempo*, la sua *localizzazione*. L'idea è che una causa influenzi l'effetto non solo in termini binari (accade/non accade) ma lungo tutto uno spettro di modi in cui l'effetto potrebbe manifestarsi. La causalità come influenza cattura allora questa variazione multidimensionale.

Formalmente, siano c ed e eventi che ammettono non solo l'alternativa binaria tra occorrenza e non-occorrenza, ma una gamma di varianti: c potrebbe verificarsi prima o dopo, con maggiore o minore intensità, in posizioni spaziali leggermente diverse; analogamente e. Lewis definisce c come causa di e se e dipende da c nel senso che variazioni nello *stile* di accadimento di c si riflettono in variazioni nello stile di accadimento di e nei mondi vicini. Questo approccio cattura casi di overdetermination simmetrica che la teoria precedente mancava, poiché le variazioni nel contributo di c_1 producono effettivamente variazioni nell'effetto (magari infinitesimali), mentre l'eventuale sostituzione con c_2 avrebbe prodotto un effetto leggermente diverso.

La complessità crescente di queste revisioni suggerisce che la struttura controfattuale , nella versione lewisiana, porta con sé un onere di aggiustamento progressivo che, pur non invalidando il programma, lo carica di impegni sempre più pesanti sul piano metafisico.

II.6 — Proprietà logiche della struttura lewisiana

È utile raccogliere, in forma sintetica, le proprietà formali della logica **VC** e dei sistemi correlati che Lewis costruisce. Sia $\mathbf{L} = \langle \mathbf{W}, \{\leq_w\}\{w \in W\}, \mathbf{V}\rangle$ un modello lewisiano. Le seguenti equivalenze e proprietà di validità caratterizzano il comportamento dell'operatore $\Box\!\!\rightarrow$:

Equivalenza con quantificazione su mondi prossimi: $\varphi \; \Box\!\!\rightarrow \psi$ è vero in $w \Leftrightarrow \forall u \in \mathbf{f}_\varphi(w) : \mathbf{V}(\psi, u) = 1$

dove $\mathbf{f}_\varphi(w) = \{u : \mathbf{V}(\varphi,u)=1 \wedge \neg \exists v(v <_w u \wedge \mathbf{V}(\varphi,v)=1)\}$ è il selettore.

Monotonia verso il basso della conseguente: se $\varphi \; \Box\!\!\rightarrow \psi$ e $\psi \vDash \chi$ (ψ implica classicamente χ), allora $\varphi \; \Box\!\!\rightarrow \chi$. Questa proprietà — spesso chiamata Right Weakening — vale perché se tutti i mondi φ-veri più vicini soddisfano ψ, e ψ implica χ in tutti i mondi, allora quei mondi soddisfano anche χ.

Non-monotonia verso l'antecedente: $\varphi \boxrightarrow \psi$ non implica $(\varphi \wedge \chi) \boxrightarrow \psi$. I mondi più vicini in cui $\varphi \wedge \chi$ vale possono differire dai mondi più vicini in cui vale φ soltanto, annullando la conseguenza.

Importazione (valida parzialmente): $(\varphi \wedge \psi) \boxrightarrow \chi \Leftrightarrow \varphi \boxrightarrow (\psi \boxrightarrow \chi)$, purché si assuma l'assioma di Centering forte e alcune condizioni sull'ordinamento. Questa proprietà — l'equivalenza tra il condizionale con antecedente congiunto e il condizionale annidato — è la via attraverso cui Lewis connette la struttura dei controfattuali itinerati.

Deducibilità e consistenza: la logica **VC** è completa rispetto alla classe dei modelli con ordinamenti totali e l'assioma di Limite. Il sistema più debole **VW** (senza l'assioma di Limite) è completo rispetto agli ordinamenti senza la condizione di minimo.

Il risultato di completezza, dimostrato da Lewis con metodologia standard della semantica delle logiche modali, assicura che ogni tautologia di **VC** è dimostrabile nel sistema assiomatico corrispondente e viceversa, collocando così la logica controfattuale lewisiana all'interno della famiglia delle logiche modali non-normali, precisamente nella varietà delle logiche di *condizionale variabilmente rigoroso* (*variably strict conditional*), espressione coniata da Lewis stesso per distinguere il suo operatore da quelli della logica modale standard.

III. Judea Pearl e i modelli causali strutturali

III.1 — Un cambio di paradigma: dalla semantica all'ingegneria della causalità

Il progetto di Judea Pearl rappresenta, rispetto a quello lewisiano, uno spostamento di prospettiva che non è soltanto tecnico ma concettuale. Laddove Lewis muove dalla logica del linguaggio ordinario — dalle condizioni di verità degli enunciati che gli esseri umani effettivamente proferiscono quando ragionano su ciò che sarebbe accaduto se le cose fossero andate diversamente — Pearl muove dai requisiti operativi della statistica e dell'apprendimento automatico: come può un agente che osserva dati tratti da un sistema fisico, biologico o sociale *inferire* la struttura causale di quel sistema? Come si distingue la mera correlazione, che un'operazione di condizionamento statistico può sempre produrre, dalla dipendenza causale genuina, che resiste alla manipolazione del condizionante?

Questa differenza di prospettiva è profonda e feconda. Per Lewis, i controfattuali sono il *linguaggio* della causalità: siamo a livello semantico, vogliamo sapere che cosa significano. Per Pearl, i controfattuali sono il *contenuto* della causalità: siamo a livello epistemico-computazionale, vogliamo sapere come derivarli da dati osservativi e strutturali. Il risultato è che Pearl approda a una teoria in cui la struttura causale è esplicita, rappresentabile, e manipolabile con strumenti algebrici, mentre Lewis rimane

in un quadro in cui la struttura causale è mediata dall'intuizione sui mondi possibili più vicini, con tutto il bagaglio di indeterminatezza che questo comporta.

III.2 — I modelli causali strutturali: architettura formale

Il cuore del sistema pearliano è la nozione di *modello causale strutturale* (SCM, *Structural Causal Model*). Un tale modello è una quadrupla:

$$\mathbf{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$$

Le componenti si articolano come segue.

U è l'insieme delle *variabili esogene* o di sfondo (*background variables*). Queste variabili rappresentano tutti i fattori che influenzano il sistema in esame ma la cui genesi causale non è modellata all'interno di **M**: sono, per così dire, i "rumori" o le condizioni al contorno che il modello riceve dall'esterno senza spiegarli. Nella pratica applicata, **U** cattura l'eterogeneità non osservata della popolazione o del sistema.

V è l'insieme delle *variabili endogene* (*endogenous variables*), quelle la cui genesi causale è modellata esplicitamente all'interno di **M**. Ogni $V_i \in \mathbf{V}$ è una variabile il cui valore è determinato, nel modello, da una funzione strutturale che dipende da altre variabili del sistema e da componenti esogene.

F è l'insieme delle *funzioni strutturali* (o *equazioni strutturali*): per ogni $V_i \in \mathbf{V}$, esiste una funzione:

$$f_i : \mathrm{dom}(PA_i) \times \mathrm{dom}(U_i) \to \mathrm{dom}(V_i)$$

dove $PA_i \subseteq \mathbf{V} \setminus \{V_i\}$ è l'insieme dei *genitori causali diretti* di V_i nel modello — le variabili endogene che influenzano direttamente V_i — e $U_i \subseteq \mathbf{U}$ è il sottoinsieme delle variabili esogene che incidono su V_i. L'equazione strutturale si scrive:

$$V_i := f_i(PA_i, U_i)$$

Il simbolo ":=" (assegnazione strutturale) è deliberatamente distinto dall'uguale delle equazioni ordinarie. Un'equazione strutturale non è una relazione simmetrica tra grandezze: afferma che il *valore* di V_i è *determinato* dai valori di PA_i e U_i secondo la funzione f_i, e non viceversa. Questa asimmetria è il cuore della causalità nella prospettiva pearliana.

P(U) è la distribuzione di probabilità congiunta sulle variabili esogene. Poiché ogni V_i è una funzione deterministica di PA_i e U_i, la randomicità del modello proviene interamente dalla distribuzione di **U**. Ne segue che $P(\mathbf{U})$ induce una distribuzione di probabilità su tutte le variabili di **V**, che è la distribuzione osservazionale $P(\mathbf{V})$ del modello.

Un modello strutturale è *ricorsivo* (o *aciclico*) se esiste un ordinamento topologico delle variabili $V_1, V_2, \ldots, V_n \in \mathbf{V}$ tale che ogni V_i dipende solo da variabili V_j con $j < i$. In un modello ricorsivo, il sistema di equazioni strutturali ha sempre una

soluzione unica per ogni configurazione di **U**, il che garantisce la coerenza del modello. La classe dei modelli ricorsivi è quella di gran lunga più trattata in letteratura, e su di essa si concentra la teoria di Pearl nella sua forma standard.

III.3 — Il grafo causale come rappresentazione visiva della struttura

A ogni modello strutturale **M** è associato in modo canonico un *grafo causale aciclico orientato* (DAG, *Directed Acyclic Graph*) **G** = (**V**, **E**), dove **V** è il medesimo insieme di variabili endogene del modello e l'insieme degli archi **E** è definito da:

$(V_j \to V_i) \in \mathbf{E} \Leftrightarrow V_j \in PA_i$

Ogni arco orientato $V_j \to V_i$ codifica il fatto che V_j è un genitore causale diretto di V_i, ovvero che f_i dipende effettivamente dal valore di V_j (e non solo formalmente). Le variabili esogene U_i non compaiono come nodi espliciti nel DAG standard — sono "assorbite" nelle variabili endogene cui si riferiscono — ma in alcune estensioni del formalismo (i cosiddetti *SWIG*, *Single World Intervention Graphs*, o i modelli con variabili latenti esplicite) vengono rappresentate esplicitamente.

Il DAG ha, in questo quadro, uno statuto doppio: è a un tempo una rappresentazione visiva del modello strutturale e uno strumento per ragionare sulle *indipendenze condizionali* che il modello implica. Questa seconda funzione è catturata dalla nozione di *d-separazione*, introdotta da Pearl insieme a Thomas Verma: due insiemi di variabili **X** e **Y** sono d-separati da un insieme di variabili **Z** nel DAG **G** — scritto $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G$ — se ogni *cammino* (sequenza alternante di archi orientati e non-orientati, ignorando l'orientamento) tra ogni nodo di **X** e ogni nodo di **Y** è "bloccato" da **Z** secondo le regole seguenti:

Un cammino è bloccato da **Z** se contiene almeno uno tra i seguenti tipi di configurazioni:

(a) Catena o forchetta bloccata: il cammino passa per un nodo $V_k \in \mathbf{Z}$ in una configurazione catena ($V_i \to V_k \to V_j$) o forchetta ($V_i \leftarrow V_k \to V_j$). Il condizionamento su V_k blocca il flusso di informazione.

(b) Collider non aperto: il cammino passa per un nodo $V_k \notin \mathbf{Z}$ e nessuno dei discendenti di V_k appartiene a **Z**, in una configurazione collider ($V_i \to V_k \leftarrow V_j$). Un nodo collider blocca naturalmente il flusso di informazione a meno che non si condizioni su di esso o su un suo discendente.

Il Teorema di Markov per DAG afferma che in un modello strutturale ricorsivo, la distribuzione indotta $P(\mathbf{V})$ soddisfa:

$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G \Rightarrow X \perp\!\!\!\perp_P Y \mid Z$

dove $\perp\!\!\!\perp_P$ denota l'indipendenza statistica nella distribuzione. La d-separazione è dunque un criterio grafico che legge le indipendenze condizionali dal DAG senza

dover eseguire calcoli probabilistici espliciti. La *fedeltà* (*faithfulness*) è la condizione aggiuntiva, spesso assunta, per cui vale anche la direzione inversa: le uniche indipendenze condizionali di P sono quelle che la d-separazione predice.

III.4 — L'operatore do: interventi come mutilazioni del grafo

La distinzione concettuale più importante nel programma di Pearl è quella tra *osservazione* e *intervento*. Osservare che una variabile X assume il valore x — condizionare su X = x — è un'operazione che altera la distribuzione di probabilità ma non la struttura causale: si restringe semplicemente lo spazio campionario alle situazioni in cui X = x, tenendo invariate le relazioni strutturali. Per contro, *intervenire* sulla variabile X impostando il suo valore a x — denotato **do**(X = x) nell'operatore introdotto da Pearl — è un'operazione che *modifica la struttura causale* del modello: si elimina l'equazione strutturale f_X che determinava il valore di X in funzione dei suoi genitori, sostituendola con la costante x. I genitori causali di X perdono la loro influenza su X nell'intervento, anche se il valore di x coincide con quello che X avrebbe assunto per natura.

Formalmente, l'intervento **do**(X = x) su un modello $\mathbf{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ produce un modello *mutilato*:

$$\mathbf{M}_x = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}_x, P(\mathbf{U}) \rangle$$

dove $\mathbf{F}_x$ differisce da **F** soltanto per la funzione relativa a X: essa viene sostituita dalla funzione costante X := x (indipendentemente da PA_X e U_X). Nel DAG corrispondente, questa mutilazione si traduce nell'eliminazione di tutti gli archi entranti nel nodo X. Il DAG mutilato $\mathbf{G}_{\bar{x}}$ si ottiene da **G** cancellando tali archi.

La distribuzione indotta dal modello mutilato su una variabile Y è la *distribuzione interventistica*:

$$P(Y = y \mid \mathbf{do}(X = x)) = P_{\{\mathbf{M}_x\}}(Y = y)$$

che si distingue, in generale, dalla distribuzione condizionale osservazionale P(Y = y | X = x). Questa distinzione è il cuore del celebre *paradosso di Simpson* che Pearl usa come fil rouge pedagogico: in un dataset osservazionale, la correlazione tra un trattamento e un esito può essere invertita o annullata quando si condiziona su una variabile confondente; l'analisi interventistica, per contro, risponde a una domanda genuinamente causale e non è soggetta a questa inversione.

III.5 — Il calcolo degli interventi: il do-calculus

Pearl dimostra che, sotto ipotesi di identificabilità, la distribuzione interventistica P(Y | **do**(X = x)) può essere calcolata a partire dalla sola distribuzione osservazionale P(**V**) mediante regole algebriche — a condizione di conoscere la struttura del DAG. Il sistema di regole che permette questo calcolo è il *do-calculus*, formalizzato da Pearl nei suoi lavori degli anni Novanta.

Siano **X**, **Y**, **Z**, **W** insiemi disgiunti di variabili in **V**. Siano **G** il DAG del modello, $\mathbf{G}\{\overline{\boldsymbol{X}}\}$ *il DAG ottenuto eliminando tutti gli archi entranti in* ***X****, e* $\boldsymbol{G}\{\underline{\mathbf{X}}\}$ il DAG ottenuto eliminando tutti gli archi uscenti da **X**. Denotando con $\mathbf{G}_{\{\overline{\mathbf{X}},\underline{\mathbf{Z}}\}}$ il DAG ottenuto eliminando gli archi entranti in **X** *e* gli archi uscenti da **Z**, si enunciano le tre regole del do-calculus:

Regola 1 — Inserimento/eliminazione di osservazioni:

$P(\mathbf{Y} \mid \mathbf{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W}) = P(\mathbf{Y} \mid \mathbf{do}(\mathbf{X}), \mathbf{W})$

se vale $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})\{\boldsymbol{G}\{\overline{\mathbf{X}}\}\}$

La regola afferma che se **Z** è d-separato da **Y** dato **X** e **W** nel DAG mutilato $\mathbf{G}_{\{\overline{\mathbf{X}}\}}$ (ovvero nel modello post-intervento su **X**), allora la conoscenza di **Z** è irrilevante per la previsione di **Y** nell'intervento su **X** condizionatamente a **W**. In sostanza, si può inserire o eliminare variabili osservate dalla condizione purché esse risultino d-separate dall'effetto nel grafo appropriato.

Regola 2 — Scambio intervento/osservazione:

$P(\mathbf{Y} \mid \mathbf{do}(\mathbf{X}), \mathbf{do}(\mathbf{Z}), \mathbf{W}) = P(\mathbf{Y} \mid \mathbf{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W})$

se vale $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})\{\boldsymbol{G}\{\overline{\mathbf{X}},\underline{\mathbf{Z}}\}\}$

dove $\mathbf{G}_{\{\overline{\mathbf{X}},\underline{\mathbf{Z}}\}}$ è il DAG con gli archi entranti in **X** *e* gli archi entranti in **Z** eliminati.

La regola afferma che un intervento su **Z** può essere sostituito da una mera osservazione di **Z** (e viceversa) quando, nel grafo con intervento su **X** e intervento su **Z**, la variabile **Z** risulta d-separata da **Y** dato **X** e **W**. Questa regola è lo strumento principale per trasformare espressioni interventistiche in espressioni osservazionali, ovvero per ridurre quantità causali a quantità statistiche.

Regola 3 — Inserimento/eliminazione di interventi:

$P(\mathbf{Y} \mid \mathbf{do}(\mathbf{X}), \mathbf{do}(\mathbf{Z}), \mathbf{W}) = P(\mathbf{Y} \mid \mathbf{do}(\mathbf{X}), \mathbf{W})$

se vale $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}, \mathbf{W})\{\boldsymbol{G}\{\overline{\mathbf{X}},\overline{\mathbf{Z}(\mathbf{W})}\}\}$

dove $\mathbf{Z}(\mathbf{W}) = \mathbf{Z} \setminus \mathrm{An}(\mathbf{W})\{\boldsymbol{G}\{\overline{\mathbf{X}}\}\}$ è il sottoinsieme di **Z** che *non* è antenato di alcun nodo di **W** nel grafo $\mathbf{G}_{\{\overline{\mathbf{X}}\}}$.

La terza regola gestisce il caso più delicato: un intervento su **Z** può essere eliminato del tutto dalla condizione quando le variabili di **Z** che non sono antenati di **W** nel grafo mutilato sono d-separate da **Y**. Questa regola è necessaria per trattare casi in cui

l'intervento su **Z** potrebbe influenzare **W**, e dunque indirettamente **Y**, attraverso percorsi che passano per **W**.

La completezza del do-calculus — ovvero il fatto che queste tre regole siano sufficienti per ricavare *tutte* le identità tra quantità interventistiche e osservazionali che possono essere ricavate dalla struttura del DAG — è stata dimostrata da Shpitser e Pearl (2006) e, indipendentemente, da Huang e Valtorta, con risultati che confermano che il do-calculus caratterizza esattamente la classe dei problemi di identificazione causale su DAG semimarkoviani.

III.6 — La formula di aggiustamento e l'aggiustamento per il backdoor

Un caso particolare di applicazione del do-calculus, che riveste importanza sia teorica che applicativa, è la *formula di aggiustamento* (*adjustment formula*). Si supponga che **Z** soddisfi il *criterio del backdoor* rispetto alla coppia (**X**, **Y**): ovvero che nessun nodo di **Z** sia discendente di **X**, e che **Z** blocchi ogni *cammino di backdoor* tra **X** e **Y** in **G** — cioè ogni cammino che inizia con una freccia entrante in **X**.

In questo caso, la distribuzione interventistica è data da:

$$P(Y = y \mid \mathbf{do}(X = x)) = \Sigma_z \, P(Y = y \mid X = x, \mathbf{Z} = z) \cdot P(\mathbf{Z} = z)$$

ovvero dalla media ponderata delle probabilità condizionali di Y dato X e **Z**, pesata con la distribuzione marginale di **Z**. Questa formula — nota anche come *formula del regressore* o *formula di standardizzazione causale* — permette di calcolare l'effetto causale di X su Y dalla sola distribuzione osservazionale, purché i confondenti siano misurati e soddisfino il criterio del backdoor.

È importante sottolineare che questa formula non è una semplice standardizzazione statistica: la sua derivazione richiede l'assunzione esplicita della struttura causale codificata nel DAG, e il risultato ha un'interpretazione causale — non meramente associativa — soltanto in virtù di quella struttura. È questo il punto in cui il formalismo pearliano rende esplicito il confine, nella pratica scientifica, tra descrizione statistica e inferenza causale.

III.7 — La scaletta causale: associazione, intervento, controfattuale

Un contributo di grande chiarezza concettuale del programma pearliano è la sistematizzazione dei diversi livelli epistemici del ragionamento causale nella cosiddetta *scala causale* (*ladder of causation*), che Pearl articola — specialmente in collaborazione con Dana Mackenzie — in tre gradini distinti.

Il primo gradino è quello dell'*associazione*: risponde a domande del tipo "Quanto spesso Y = y quando X = x?", cioè a domande sulla distribuzione condizionale P(Y | X).

Questo livello è interamente catturato dalla statistica classica e non richiede alcuna ipotesi causale. Un agente in grado soltanto di osservare e correlare, ma non di intervenire o ipotizzare, si ferma a questo piano.

Il secondo gradino è quello dell'*intervento*: risponde a domande del tipo "Se ponessimo X = x, come cambierebbe Y?", cioè a domande sulla distribuzione interventistica P(Y | **do**(X = x)). Questo livello richiede la conoscenza della struttura causale del sistema — il DAG e le equazioni strutturali — e non può essere ridotto al primo gradino senza ulteriori assunzioni. Un agente in grado di intervenire fisicamente sul sistema, ma non di ragionare su scenari ipotetici non realizzati, si colloca a questo piano.

Il terzo gradino è quello del *controfattuale*: risponde a domande del tipo "Dato che X = x' e Y = y' si sono verificati, cosa sarebbe accaduto a Y se X avesse assunto il valore x?", ovvero a domande sulla distribuzione controfattuale P(Y_x = y | X = x', Y = y'). Questo livello è il più ricco e il più esigente: richiede non solo la struttura del DAG ma anche la specificazione del meccanismo di generazione, inclusa la distribuzione delle variabili esogene o almeno la distribuzione congiunta delle variabili osservate condizionate alle evidenze.

L'innovazione di Pearl consiste nel mostrare che il terzo gradino è, in linea di principio, formalmente *derivabile* dalla struttura del modello strutturale, e che i tre livelli formano una gerarchia strettamente crescente: ciò che è derivabile al secondo livello è derivabile anche al terzo, ma non viceversa; ciò che è derivabile al primo è derivabile anche al secondo, ma non viceversa. Questa gerarchia formalizza, in modo preciso, l'intuizione comune che la causalità sia "qualcosa di più" della correlazione.

III.8 — I controfattuali nel quadro strutturale: la procedura ABP

La derivazione formale dei controfattuali nel quadro pearliano si realizza mediante la procedura *Abduction-Action-Prediction* (ABP), che si articola in tre fasi sequenziali.

Fase 1 — Abduzione (*Abduction*): si condiziona la distribuzione delle variabili esogene P(**U**) all'evidenza osservata **e** = {**V** = **v**}, ottenendo la distribuzione aggiornata:

P(**U** | **e**) = P(**U** | **V** = **v**)

Questa fase è un'operazione bayesiana standard: si usano le equazioni strutturali per risalire dai valori osservati delle variabili endogene ai valori più probabili (o alla distribuzione più probabile) delle variabili esogene che li hanno generati. In un modello deterministico — in cui ogni U_i determina univocamente V_i dati PA_i — e con osservazione completa di tutte le variabili endogene, questa fase produce una distribuzione degenere, ovvero un singolo valore di **u** compatibile con le osservazioni.

Fase 2 — Azione (*Action*): si applica l'intervento **do**(X = x) al modello, producendo il modello mutilato **M**_x nel modo già descritto. Il modello mutilato mantiene la distribuzione aggiornata P(**U** | **e**) sulle variabili esogene, ma le equazioni strutturali di X vengono sostituite dalla costante.

Fase 3 — Previsione (*Prediction*): si calcola la distribuzione di Y nel modello mutilato **M**_x con le variabili esogene distribuite secondo P(**U** | **e**):

P(Y_x = y | **e**) = P_{**M**_x}(Y = y | P(**U**) = P(**U** | **e**))

Il risultato è la distribuzione controfattuale di Y sotto l'ipotesi che X fosse stato impostato a x, condizionata sull'evidenza effettivamente osservata **e**. Questa distribuzione risponde a domande del tipo: "Dato che ho osservato questi dati, cosa sarebbe accaduto a Y se X fosse stato x?".

Nel caso di un individuo specifico — una *unità* identificata da un valore **u**_0 del vettore esogeno — il controfattuale è determinato in modo puntuale:

Y_x(**u**_0) = valore di Y nel modello **M**_x con **U** = **u**_0

Questa quantità è il *controfattuale individuale* o *controfattuale di unità*, che Pearl usa sistematicamente per definire nozioni come il *Nesso Causale Necessario* (NCN) e il *Nesso Causale Sufficiente* (NCS):

NCN: P(Y_{x'}(**u**) ≠ y' | X = x, Y = y) — la probabilità che, in assenza del trattamento, l'esito non si sarebbe verificato

NCS: P(Y_x(**u**) = y | X = x', Y = y') — la probabilità che, in presenza del trattamento, l'esito si sarebbe verificato comunque

Questi due indici — insieme al *Nesso Causale Necessario e Sufficiente* (NCNS) — costituiscono la formalizzazione pearliana del concetto giuridico e quotidiano di "causa prossima", "causa sine qua non" e "causa ma-for", rispettivamente, aprendo un ponte diretto tra la teoria formale e le applicazioni in campo medico-legale, epidemiologico e politico.

III.9 — Il modello gemello e la semantica dei controfattuali

Una rappresentazione alternativa — e spesso più intuitiva — della procedura ABP è il metodo del *modello gemello* (*twin network*), che permette di visualizzare la struttura controfattuale direttamente in termini grafici.

Il modello gemello **M**^{(x)} associato a **M** e all'intervento **do**(X = x) consiste di due copie del DAG originale, che condividono le variabili esogene **U** ma hanno variabili endogene distinte. La prima copia — il *mondo fattuale* — rappresenta il modello originale **M** con le variabili endogene V_1, V_2, …, V_n che soddisfano le equazioni strutturali originali. La seconda copia — il *mondo controfattuale* — rappresenta il modello mutilato **M**_x con le variabili endogene V_1^{(x)}, V_2^{(x)}, …, V_n^{(x)}, di cui X^{(x)} = x per costruzione.

Il fatto cruciale è che le variabili esogene **U** sono condivise tra le due copie: il contesto (**U** = **u**) è lo stesso nel mondo fattuale e nel mondo controfattuale. Questo cattura l'intuizione che il controfattuale riguardi *lo stesso individuo* o *la stessa unità* — identifica

per valori di **u** — in due situazioni ipotetiche distinte. Nel modello gemello, la distribuzione congiunta di quantità fattuali e controfattuali è ben definita, e permette di calcolare probabilità di enunciati come P(Y_x = y ∧ Y_{x'} = y'), fondamentali per definire il nesso causale necessario e sufficiente.

La struttura del modello gemello ricorda, in modo non casuale, la struttura dei *mondi possibili* di Lewis. In entrambi i casi, si considerano due situazioni — il mondo in cui l'antecedente è vero e il mondo fattuale — e si chiede qual è la relazione tra le conseguenze nelle due situazioni. La differenza chiave è che nel modello gemello la connessione tra i due "mondi" è esplicitata attraverso le equazioni strutturali e le variabili esogene condivise, mentre in Lewis la connessione è catturata dall'ordinamento di somiglianza, che è una relazione primitiva non ulteriormente analizzata. Il modello gemello è, in un certo senso, il *meccanismo* che Lewis non specifica.

IV. Il calcolo degli interventi: identificabilità, strutture semi-markoviane e l'algoritmo ID

IV.1 — Il problema dell'identificazione causale

Il do-calculus, nella sua formulazione delle tre regole, è uno strumento di *trasformazione* algebrica: permette di riscrivere espressioni che contengono l'operatore **do** in espressioni che ne sono prive, purché esistano opportune condizioni di d-separazione nei grafi rilevanti. Ma questo strumento non risponde ancora a una domanda preliminare e fondamentale: dato un DAG **G** e una query causale Q della forma P(**Y** | **do**(**X**), **Z**), è possibile in linea di principio esprimere Q come funzione della sola distribuzione osservazionale P(**V**)? E se sì, tale espressione è unica?

Questo è il *problema dell'identificazione causale*. La sua importanza pratica è immensa: nella stragrande maggioranza delle situazioni di ricerca empirica, lo scienziato non ha la possibilità di eseguire interventi controllati (esperimenti randomizzati) e deve accontentarsi di dati osservativi. Se la quantità causale di interesse è identificabile dalla distribuzione osservazionale, allora si può rispondere a domande causali senza esperimenti; se non lo è, occorre raccogliere dati sperimentali o rassegnarsi all'ignoranza.

La distinzione tra DAG *markoviani* — in cui tutte le variabili causalmente rilevanti sono osservate — e DAG *semi-markoviani* — in cui alcune variabili esogene influenzano più variabili endogene, creando così *confondenti latenti* — è cruciale a questo riguardo. Nei DAG markoviani, sotto le ipotesi di Markov e fedeltà, l'effetto causale di **X** su **Y** è sempre identificabile quando esiste almeno un insieme **Z** che soddisfa il criterio del backdoor. Ma quando sono presenti confondenti latenti — rappresentati graficamente

da archi bidirezionali tra le variabili endogene (archi tratteggiati, nella convenzione di Pearl) — l'identificabilità non è più garantita e deve essere verificata caso per caso.

IV.2 — Criteri di aggiustamento: backdoor e frontdoor

Il *criterio del backdoor* è il risultato di identificabilità più direttamente intuitivo e storicamente primo. Un insieme di variabili **Z** soddisfa tale criterio rispetto alla coppia ordinata (**X**, **Y**) nel DAG **G** se:

(i) nessun nodo di **Z** è discendente di **X** in **G**;

(ii) **Z** blocca ogni cammino di backdoor tra **X** e **Y** in **G**, dove un cammino di backdoor è qualsiasi cammino che inizia con una freccia orientata *verso* **X**.

Sotto queste due condizioni, la formula di aggiustamento vale:

$$P(Y = y \mid \mathbf{do}(X = x)) = \Sigma_{\mathbf{z}} P(Y = y \mid X = x, \mathbf{Z} = \mathbf{z}) \cdot P(\mathbf{Z} = \mathbf{z})$$

La condizione (i) è necessaria perché condizionare su un discendente di X può aprire cammini di backdoor che erano bloccati (un fenomeno noto come *collider bias* o *M-bias*), distorcendo il risultato. La condizione (ii) assicura che tutti i percorsi non-causali tra X e Y — quelli che sfruttano cause comuni non sul cammino causale di interesse — vengano controllati.

Il *criterio del frontdoor* affronta una situazione complementare: quella in cui i confondenti non osservati tra X e Y rendono inapplicabile qualunque insieme di backdoor, ma esiste una variabile mediatrice **M** il cui effetto su **Y** è identificabile. L'enunciazione è la seguente: un insieme **M** soddisfa il criterio del frontdoor rispetto a (**X**, **Y**) se:

(i) tutte le vie causali da **X** a **Y** sono intercettate da **M** (ovvero **M** media *tutti* gli effetti di X su Y);

(ii) non vi sono cammini di backdoor non bloccati da **X** a **M** (ovvero l'effetto di X su M è identificabile);

(iii) tutti i cammini di backdoor da **M** a **Y** sono bloccati da **X**.

In questo caso, l'effetto causale è calcolabile mediante la *formula del frontdoor*:

$$P(Y = y \mid \mathbf{do}(X = x)) = \Sigma_{\mathbf{m}} \left[\Sigma_{x'} P(Y = y \mid X = x', \mathbf{M} = \mathbf{m}) \cdot P(X = x') \right] \cdot P(\mathbf{M} = \mathbf{m} \mid X = x)$$

La derivazione di questa formula costituisce uno dei risultati più eleganti del do-calculus. Si procede in due passi: prima si identifica $P(\mathbf{M} \mid \mathbf{do}(X)) = P(\mathbf{M} \mid X)$ usando il fatto che non vi sono cammini di backdoor aperti da X a **M**; poi si identifica $P(Y \mid \mathbf{do}(\mathbf{M})) = \Sigma_{x'} P(Y \mid X = x', \mathbf{M}) \cdot P(X = x')$ usando X come variabile di aggiustamento per il backdoor da **M** a Y. Combinando i due passi tramite la legge della probabilità totale applicata all'intervento su **M**, si ottiene la formula sopra enunciata.

L'esempio paradigmatico del criterio del frontdoor è lo studio del fumo e del cancro ai polmoni in presenza di un ipotetico gene confondente: non si possono assegnare casualmente i soggetti al fumo, e il gene confondente non è osservato. Tuttavia, se si misura la presenza di catrame nei polmoni come mediatore sul cammino causale, e se le condizioni del criterio sono soddisfatte, l'effetto causale del fumo sul cancro è identificabile attraverso la formula del frontdoor.

IV.3 — Strutture semi-markoviane e confondenti latenti

Quando il DAG contiene variabili latenti — variabili non osservate che influenzano più di una variabile endogena osservata — la rappresentazione grafica più adeguata è il *CADMG* (*Complete Acyclic Directed Mixed Graph*) o, nella terminologia di Pearl, il grafo semi-markoviano, in cui gli archi bidirezionali ↔ (o archi tratteggiati curvi) denotano l'esistenza di un confondente latente comune tra le due variabili collegate.

Formalmente, un arco bidirezionale V_i ↔ V_j nel grafo semi-markoviano abbrevia l'esistenza di una variabile esogena U_{ij} ∈ **U** che compare nelle equazioni strutturali sia di V_i che di V_j, senza essere osservata. La distribuzione osservazionale è quindi una *proiezione marginale* della distribuzione del modello completo sulle sole variabili di **V**.

In presenza di confondenti latenti, l'identificabilità può venire a mancare. L'esempio canonico di non-identificabilità è il caso di un arco bidirezionale diretto tra X e Y:

X ↔ Y

In questo caso, qualunque effetto causale di X su Y è confuso con la correlazione indotta dal confondente latente, e nessun insieme di backdoor è disponibile (poiché il confondente latente non è misurabile). La formula di aggiustamento non si applica, e il do-calculus non produce alcuna espressione puramente osservazionale per P(Y | **do**(X)). L'effetto causale è *non identificabile* da dati osservativi soli.

Il risultato di non-identificabilità ha una caratterizzazione grafica precisa nel criterio dell'*hedge* (*siepe*), introdotto da Shpitser e Pearl: una quantità causale P(**Y** | **do**(**X**)) non è identificabile in un grafo semi-markoviano se e soltanto se nel grafo esiste una struttura chiamata *hedge*, composta da due sottografi **F** e **F'** che soddisfano determinate condizioni di raggiungibilità e connessione. La dimostrazione della completezza di questo criterio — che ogni quantità non identificabile corrisponde a un hedge nel grafo — è il risultato centrale dell'articolo di Shpitser e Pearl del 2006 e costituisce, assieme alla completezza del do-calculus per le quantità identificabili, la caratterizzazione definitiva del problema di identificazione causale su DAG semi-markoviani.

IV.4 — L'algoritmo ID: identificazione sistematica

L'algoritmo ID (*Identification Algorithm*), sviluppato da Ilya Shpitser e Judea Pearl, è una procedura ricorsiva che, dato un DAG semi-markoviano **G** e una query causale P(**y** |

do(**x**)), determina in tempo finito se la query è identificabile e, in caso affermativo, restituisce la formula esplicita in termini di P(**V**).

La struttura dell'algoritmo si fonda su sette regole di riscrittura, di cui le prime tre corrispondono ai casi base e le restanti quattro gestiscono la ricorsione su sottografi. L'idea di fondo è di decomporre progressivamente la query causale in componenti più semplici, applicando a ciascuna le trasformazioni appropriate fino a ottenere espressioni puramente osservazionali.

Denotiamo con **ID**(**y**, **x**, P, **G**) la chiamata principale dell'algoritmo, che restituisce un'espressione per P(**y** | **do**(**x**)) nel modello corrispondente a **G**, data la distribuzione P.

La struttura ricorsiva si sviluppa attraverso i seguenti casi principali (presentati in forma semplificata ma fedele alla logica dell'algoritmo):

Caso 1: se **x** = ∅, restituire Σ_{**v** \ **y**} P(**v**). Nessun intervento significa che si tratta di una semplice marginale della distribuzione osservazionale.

Caso 2: se **V** = An(**Y**)***G*** *(l'insieme di tutte le variabili è uguale all'insieme degli antenati di **Y** in **G**), restituire* Σ{**v** \ **y**} ∏{*V_i* ∈ ***V***} *P(V_i* | *V*{π}^{(i-1)}}), dove V_{π}^{(i-1)} denota le variabili che precedono V_i in un ordinamento topologico π. In altri termini, si fattorizza la distribuzione secondo l'ordinamento topologico, eliminando le variabili non pertinenti.

Caso 3: se esistono variabili W = (**V** \ **X**) \ An(**Y**)_{**G_x̄**} — cioè variabili non in **X** e non antenati di **Y** nel grafo con gli archi entranti in **X** rimossi — allora restituire **ID**(**y**, **x** ∪ **w**, P, **G**). Le variabili che non sono antenati causali di **Y** nell'intervento su **X** sono irrilevanti per la query e possono essere aggiunte all'insieme degli interventi (rendendo la formula potenzialmente più semplice).

Caso 4: sia **CC**(**G** \ **X**) = {**S**_1, …, **S***k*} *la partizione in componenti connesse del grafo **G** \ **X** (il grafo privato dei nodi di **X**). Restituire* Σ{**v** \ (**y** ∪ **x**)} ∏_i **ID**(**s**_i, **v** \ **s**_i, P, **G**). La deconnessione del grafo permette di fattorizzare la query in sotto-problemi indipendenti.

Caso 5: se esiste S ∈ **CC**(**G** \ **X**) tale che **Y** ⊆ **S**, controllare se il grafo contiene un hedge. Se lo contiene, l'algoritmo termina con **FAIL** (la query non è identificabile). Altrimenti procedere con la ricorsione.

Caso 6: se (**S**) ∈ **CC**(**G**) (S è una componente connessa dell'intero grafo **G**, non solo di **G** \ **X**), restituire Σ_{**s** \ **y**} ∏{*V_i* ∈ ***S***} *P(V_i* | *V*{π}^{(i-1)}} ∩ **S**, V_{π}^{(i-1)} \ **S**). In questo caso la componente S è già isolata nel grafo originale e si può procedere con la fattorizzazione interna.

Caso 7 (caso generale ricorsivo): trovare la componente **S'** ∈ **CC**(**G**) tale che **S** ⊊ **S'**. Restituire:

Σ_{**s'** \ **y**} ∏{*V_i* ∈ ***S'***} *P(V_i* | *V*{π}^{(i-1)}} ∩ **S'**, V_{π}^{(i-1)} \ **S'**)

modificata in accordo con i nodi di $\mathbf{X} \cap \mathbf{S'}$ che agiscono come interventi sulla componente ristretta.

La correttezza dell'algoritmo ID — ossia il fatto che ogni formula da esso prodotta è effettivamente uguale a $P(\mathbf{y} \mid \mathbf{do}(\mathbf{x}))$ nel modello — è dimostrata per induzione sulla struttura ricorsiva. La completezza — ossia il fatto che ogni query identificabile è effettivamente calcolata dall'algoritmo — segue dal criterio dell'hedge: se l'algoritmo non produce **FAIL**, non esiste alcun hedge e la quantità è identificabile; se produce **FAIL**, l'hedge garantisce che esistono due modelli con la stessa distribuzione osservazionale ma effetti causali diversi, rendendo la quantità non identificabile.

IV.5 — Identificabilità parziale e bounds causali

Quando l'effetto causale non è identificabile in senso puntuale, il ragionamento causale non è necessariamente condannato al silenzio. Sotto ipotesi più deboli — in particolare quando le variabili sono dicotomiche e si dispone di informazioni parziali sulla struttura — è possibile derivare *limiti* (*bounds*) sull'effetto causale che restringono l'insieme dei valori compatibili con i dati osservazionali e la struttura del modello.

I *natural bounds* di Balke e Pearl per variabili binarie in presenza di un singolo confondente latente tra X e Y si derivano attraverso un programma di ottimizzazione lineare. Sia $X \in \{0,1\}$ e $Y \in \{0,1\}$. Sia $\mathbf{r}$ il vettore delle probabilità osservazionali $P(Y = y, X = x)$ per $y, x \in \{0,1\}$, a cui si aggiungono le probabilità relative a eventuali strumenti o mediatori. L'effetto causale medio (*Average Causal Effect*, ACE) è definito come:

$$\text{ACE} = P(Y = 1 \mid \mathbf{do}(X = 1)) - P(Y = 1 \mid \mathbf{do}(X = 0))$$

Nei modelli non identificabili, ACE può assumere qualsiasi valore in un intervallo $[\text{ACE}_{\text{lower}}, \text{ACE}_{\text{upper}}]$ compatibile con $\mathbf{r}$. Questo intervallo si calcola massimizzando e minimizzando ACE soggetto ai vincoli che le probabilità latenti siano non-negative e sommino a 1. Il risultato è:

$$\text{ACE}_{\text{lower}} = \max\{\, P(Y=1, X=1) + P(Y=0, X=0) - 1,\, P(Y=1) - P(X=1 \cdot Y=1) - \ldots \,\}$$

L'espressione esatta dipende dalla struttura del grafo e non può essere ridotta a una formula universale per tutti i casi; tuttavia la metodologia del programma lineare è applicabile in modo algoritmico a qualunque grafo finito con variabili discrete, e produce bounds stretti — non ulteriormente restringibili senza assunzioni aggiuntive — che quantificano esattamente il residuo di incertezza causale irriducibile dai soli dati osservativi.

IV.6 — Effetti causali condizionali e trasportabilità

Un'estensione importante del framework di identificazione riguarda il problema della *trasportabilità* (*transportability*): è possibile trasferire conclusioni causali apprese in un contesto (una popolazione, un ambiente, un tempo) a un contesto diverso? Questo

problema sorge ogni volta che le condizioni sperimentali differiscono da quelle in cui si vogliono applicare i risultati — una circostanza praticamente universale in epidemiologia, psicologia e scienze sociali.

Pearl e Bareinboim hanno formalizzato questo problema attraverso il concetto di *differenza di selezione* (*selection diagram*): un grafo aumentato in cui nodi speciali S_i denotano le differenze sistematiche tra la popolazione sorgente (in cui si dispone di dati sperimentali o causali) e la popolazione bersaglio (in cui si vogliono applicare i risultati). Un effetto causale è *trasportabile* dalla sorgente al bersaglio se è possibile riscriverlo in termini di quantità disponibili nella sorgente e nella bersaglio attraverso un'applicazione opportuna del do-calculus esteso ai nodi di selezione.

Formalmente, la query di trasportabilità è:

$P^*(Y \mid \mathbf{do}(X)) = ?$ in funzione di $P(\cdot)$, $P^*(\cdot)$, e i nodi S

dove P indica la distribuzione nella sorgente, P* quella nella bersaglio, e i nodi S codificano le differenze strutturali. La risposta richiede un algoritmo di trasportabilità — anch'esso sviluppato da Bareinboim e Pearl — che combina il do-calculus con criteri grafici sui nodi di selezione.

Il problema della trasportabilità generalizza sia l'identificabilità classica sia l'analisi della *validità esterna* degli studi randomizzati: dimostrare che un effetto causale trovato in un trial clinico vale anche per la popolazione generale è, in questo quadro, un problema di trasportabilità, e la risposta dipende dalla struttura delle differenze tra la popolazione del trial e quella generale.

IV.7 — Variabili continue e equazioni strutturali lineari

Sebbene il framework pearliano sia presentato spesso con variabili discrete per semplicità espositiva, la sua applicazione più classica nella letteratura econometrica e statistica riguarda variabili continue con equazioni strutturali lineari. In questo caso, il modello strutturale assume la forma:

$$V_i = \Sigma_{V_j \in PA_i} \beta_{ji} \cdot V_j + \varepsilon_i$$

dove β_{ji} sono i *coefficienti strutturali* (*structural coefficients*) — da distinguersi dai coefficienti di regressione ordinari — e ε_i sono termini di errore con media zero, potenzialmente correlati tra loro (i confondenti latenti corrispondono a correlazioni non zero tra diversi ε_i).

In un sistema lineare ricorsivo senza confondenti latenti, la matrice dei coefficienti strutturali $\mathbf{B}$ (con $B_{ij} = \beta_{ji}$) è triangolare inferiore (per il carattere aciclico del DAG), e la distribuzione indotta sulle variabili osservate è una normale multivariata con matrice di covarianza:

$$\mathbf{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1} \, \mathbf{\Psi} \, [(\mathbf{I} - \mathbf{B})^{-1}]^T$$

dove $\mathbf{\Psi}$ è la matrice di covarianza degli errori ε_i (diagonale in assenza di confondenti latenti). L'effetto causale di X su Y — l'effetto medio dell'intervento $\mathbf{do}(X = x + \delta)$ rispetto a $\mathbf{do}(X = x)$ — è dato dal coefficiente del *cammino causale totale* da X a Y, che si calcola sommando i prodotti dei coefficienti strutturali lungo tutti i cammini orientati da X a Y nel DAG:

$$\partial P(Y \mid \mathbf{do}(X = x))/\partial x = \Sigma_{\text{cammini p da X a Y}} \prod_{\text{archi } (V_i \to V_j) \in p} \beta_{ji}$$

Questa formula — nota come *regola del cammino* (*path rule*) di Wright — è l'antecedente storico dell'intera teoria causale strutturale e risale al genetista Sewall Wright (1921), che la introdusse nell'analisi della trasmissione ereditaria dei caratteri fenotipici. Pearl ha recuperato e formalizzato questa tradizione, mostrandone la coerenza con il do-calculus nel caso lineare e generalizzando l'approccio al caso non lineare attraverso gli operatori di intervento.

Nel caso non lineare, l'effetto causale locale di X su Y nell'intorno del valore $x = x_0$ può essere approssimato dal derivato parziale dell'effetto interventistico:

$$\text{ACE}(x_0 + \delta, x_0) \approx \delta \cdot [\partial/\partial x \, P(Y \mid \mathbf{do}(X = x))]_{x = x_0}$$

ma per effetti globali — dove il sistema è fortemente non lineare — occorre ricorrere direttamente alla formula di aggiustamento o all'algoritmo ID applicato al modello non lineare.

IV.8 — La connessione con la teoria del controllo e gli interventi dinamici

Il quadro degli interventi di Pearl è concettualmente contiguo alla teoria del controllo automatico, nella misura in cui entrambe le discipline si interrogano su come modificare il comportamento di un sistema dinamico attraverso azioni mirate su alcune sue variabili. Tuttavia, la teoria di Pearl opera tipicamente su modelli statici o, al più, su modelli con struttura temporale discreta, mentre la teoria del controllo tratta sistemi dinamici a tempo continuo descritti da equazioni differenziali.

L'estensione del framework causale al dominio dinamico ha prodotto i *modelli causali strutturali a tempo continuo* e le *equazioni strutturali differenziali*, in cui l'equazione strutturale per una variabile V_i assume la forma:

$$dV_i/dt = f_i(PA_i(t), U_i(t))$$

e l'intervento $\mathbf{do}(V_i(t_0) = x)$ corrisponde a un'iniezione esogena al tempo t_0 che sovrascrive il valore di V_i senza alterare le equazioni strutturali degli altri nodi. In questo contesto, la procedura ABP si generalizza in modo naturale: l'abduzione ricava la distribuzione degli stati iniziali compatibili con le traiettorie osservate fino a t_0; l'azione modifica il valore di V_i al tempo t_0; la previsione propaghts la dinamica del sistema mutilato verso il futuro.

Questo quadro, ancora in sviluppo attivo nella letteratura, connette la teoria causale strutturale con la *teoria dell'inferenza causale nei processi stocastici*, aperta dalla scuola di Granger e formalizzata più recentemente da Mooij, Peters e collaboratori attraverso i *modelli causali strutturali a tempo discreto* e la nozione di *invarianza causale* — l'idea che le relazioni strutturali rimangano stabili attraverso diversi ambienti o interventi, caratteristica che distingue le relazioni causali dalle mere correlazioni statistiche.

V. Confronto critico, problemi aperti e prospettive

V.1 — La convergenza strutturale tra Lewis e Pearl

Nonostante le differenze di vocabolario, di motivazione e di architettura formale, vi è tra la teoria lewisiana e quella pearliana una convergenza strutturale profonda che non può essere frutto di coincidenza: entrambe riconoscono nel controfattuale il veicolo semantico privilegiato della causalità, e in entrambe la struttura logica di quel veicolo è quella di una quantificazione ristretta su uno spazio di situazioni alternative.

Nella terminologia di Lewis, la verità di $\varphi \mathrel{\square\!\!\rightarrow} \psi$ in un mondo w equivale alla verità di ψ in tutti i mondi φ-veri più vicini a w. Nella terminologia di Pearl, la verità di $Y_x = y$ dato il contesto $\mathbf{u}$ equivale al fatto che il valore di Y nel modello mutilato $\mathbf{M}_x$ con variabili esogene $\mathbf{U} = \mathbf{u}$ è y. In entrambi i casi, si considera una situazione alternativa — un "mondo diverso" — in cui l'antecedente dell'ipotesi controfattuale è vero, e si chiede cosa vi accade alla variabile conseguente.

La corrispondenza formale può essere resa precisa. Sia $\mathbf{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ un modello strutturale pearliano. A ogni valore $\mathbf{u}$ del vettore esogeno corrisponde un *contesto*, che determina in modo univoco i valori di tutte le variabili endogene attraverso la soluzione del sistema di equazioni strutturali. Ogni contesto $\mathbf{u}$ può essere interpretato come un *mondo possibile* nel senso lewisiano: specifica completamente lo stato del sistema. L'insieme di tutti i contesti — lo spazio dei valori di $\mathbf{U}$ — è, in questa lettura, lo spazio dei mondi possibili del modello.

Cosa corrisponde, in questo quadro, all'ordinamento di somiglianza lewisiano? L'intervento $\mathbf{do}(X = x)$ seleziona il mondo controfattuale modificando l'equazione strutturale di X, ma mantiene invariate tutte le equazioni delle altre variabili e, crucialmente, il valore $\mathbf{u}$ delle variabili esogene. In altri termini, il "mondo più vicino in cui $X = x$" è il mondo che differisce dal mondo attuale $\mathbf{u}$ soltanto per il valore di X — e per tutte le sue conseguenze causali, che cambiano in cascata secondo le equazioni strutturali — ma che conserva invariato il contesto di fondo $\mathbf{u}$. La *somiglianza* pearliana è dunque una somiglianza che privilegia sopra ogni cosa la conservazione del contesto esogeno, e questo corrisponde, in modo non banale, alla prima priorità lewisiana di non introdurre "grandi miracoli" — deviazioni massicce dalle leggi di natura. Il modello strutturale specifica *quali* variabili rimangono invariate nell'intervento (quelle esogene e quelle non a valle di X nel DAG) e *quali* cambiano (quelle a valle di X),

determinando così in modo meccanico l'ordinamento di somiglianza che Lewis lascia primitivo.

Questa connessione non è soltanto filosoficamente illuminante: ha anche conseguenze formali precise. Briggs (2012) e Galles e Pearl (1998) hanno mostrato che i controfattuali derivabili in un modello strutturale soddisfano gli assiomi di Lewis per i sistemi di sfere — con la riserva che la semantica strutturale è più granulare, poiché distingue tra interventi su variabili diverse che potrebbero produrre lo stesso valore dell'antecedente per vie diverse. Su questo punto si tornerà tra breve.

V.2 — La divergenza fondamentale: primitivi e ontologia

Nonostante la convergenza appena descritta, le due teorie divergono su questioni che non sono di mera tecnicità formale ma investono l'ontologia del causale e il suo rapporto con la fisica del mondo.

La prima divergenza riguarda lo statuto dei *primitivi*. Lewis assume come primitivo l'ordinamento di somiglianza tra mondi possibili e ricava da esso le condizioni di verità dei controfattuali. Pearl assume come primitivi le equazioni strutturali e ricava da esse sia i controfattuali sia — implicitamente — l'ordinamento di somiglianza. Dal punto di vista esplicativo, la teoria pearliana è quindi più profonda: invece di postulare che certi mondi siano "più vicini" di altri, spiega *perché* lo sono in termini della struttura fisica del sistema. Tuttavia, questa profondità ha un costo: le equazioni strutturali devono esse stesse essere giustificate, e la domanda "da dove vengono le equazioni strutturali?" rimanda a una teoria della struttura del mondo che Pearl non sempre affronta esplicitamente.

La seconda divergenza riguarda la *granularità degli eventi*. Lewis lavora con eventi intesi come oggetti particolari localizzati spaziotemporalmente, per cui ha senso chiedersi non solo se c si è verificato o meno, ma anche come, quando, con quale intensità. La semantica dei mondi possibili è sufficientemente ricca da rappresentare questi dettagli. Pearl lavora con *variabili* e i loro *valori*, il che è una rappresentazione granulare e discreta del reale, adeguata per molti problemi applicati ma che perde per definizione informazioni sulla struttura continua degli eventi. La revisione lewisiana del 2000 — la causalità come influenza — è, in una certa misura, un tentativo di incorporare questa granularità nella teoria controfattuale; ma essa rimane a livello di eventi, non di equazioni, e il collegamento con il formalismo pearliano richiede un lavoro di traduzione non banale.

La terza divergenza riguarda il trattamento delle *leggi di natura*. Per Lewis, le leggi di natura sono essenziali per definire l'ordinamento di somiglianza: i mondi che violano le leggi sono meno simili al mondo attuale di quelli che le rispettano, e la gerarchia delle priorità nella somiglianza è costruita proprio in funzione del rispetto o violazione di tali leggi. Per Pearl, le leggi di natura sono incorporate nelle equazioni strutturali: la relazione V_i := f_i(PA_i, U_i) è, almeno in linea di principio, un'equazione fisica o biologica che descrive il comportamento del sistema. Ma Pearl non richiede che le equazioni strutturali siano *leggi* in senso nomologico: possono essere relazioni

contingenti, meccanismi evolutisi storicamente, dispositivi ingegneristici. Questa flessibilità è un pregio applicativo ma rende la connessione tra causalità e nomologia meno esplicita che in Lewis.

V.3 — Il problema della preemption nel quadro strutturale

La preemption — lo scenario in cui una causa effettiva batte sul tempo una causa potenziale impedendole di manifestarsi — è uno dei problemi più tenaci per la teoria controfattuale. Come si è visto, Lewis vi risponde attraverso la chiusura transitiva della dipendenza controfattuale o, nella revisione del 2000, attraverso la causalità come influenza. Come si comporta il formalismo pearliano di fronte a questo scenario?

Si consideri uno schema di preemption tipica: c_1 e c_2 sono entrambi potenziali cause di e; c_1 si verifica prima, produce e, e nel frattempo disattiva c_2 prima che questa possa produrre effetti. Nel modello strutturale, questo scenario richiede una variabile D che rappresenta la disattivazione di c_2 da parte del processo innescato da c_1. Il modello ha la forma:

E := f(C_1, D) con D := g(C_1) e C_2 "bloccato" da D

L'equazione strutturale di E dipende esplicitamente da C_1 (direttamente) e da D (che a sua volta dipende da C_1). Il controfattuale $Y_{C_1 = 0}$ — il valore di E in assenza di c_1 — si calcola mutilando l'equazione di C_1 e propagando il cambiamento: con C_1 = 0, D cambia in g(0), e E cambia di conseguenza. Se il modello è costruito in modo da incorporare il fatto che C_2 avrebbe prodotto E in assenza di C_1, allora $E_{C_1 = 0} = 1$ (l'effetto si sarebbe prodotto comunque, attraverso C_2). In questo caso, il controfattuale $\neg O(C_1) \mathrel{\Box\!\!\rightarrow} \neg O(E)$ è falso, e la dipendenza diretta di E da C_1 non è catturata.

Questa è esattamente la difficoltà che afflisse Lewis. Nel quadro pearliano, la soluzione è rappresentare la struttura causale in modo sufficientemente granulare da distinguere il contributo effettivo di C_1 (attraverso il cammino $C_1 \rightarrow D \rightarrow E$ e il cammino $C_1 \rightarrow E$ diretto) da quello controfattuale di C_2. La causalità di C_1 è allora catturata non dalla dipendenza controfattuale di E rispetto a C_1, ma dalla dipendenza del *processo* specifico che ha prodotto E: l'approccio dei *cammini causali specifici* (*actual causal paths*) di Hitchcock (2001) e Woodward (2003) può essere formalizzato nel quadro strutturale attraverso il concetto di *effetto causale controllato* (*controlled direct effect*) o *effetto di cammino specifico* (*path-specific effect*).

L'*effetto diretto controllato* (*Controlled Direct Effect*, CDE) di X su Y lungo il cammino diretto (non mediato da M) è definito come:

$$\mathrm{CDE}(x, x', m) = E[Y_{x,M=m}] - E[Y_{x',M=m}]$$

dove $Y_{x,M=m}$ è il valore controfattuale di Y quando X è impostato a x e M è bloccato al valore m indipendentemente da come varierebbe naturalmente. Questo operatore — che combina un intervento su X con un blocco artificiale della mediazione — cattura l'effetto causale diretto di X su Y che *non* passa attraverso M. La

formalizzazione degli effetti di cammino specifico consente, in linea di principio, di identificare quale cammino causale è stato "attivo" in un dato contesto, risolvendo il problema della preemption a livello di rappresentazione strutturale.

V.4 — Causalità effettiva versus causalità tipo

Una distinzione fondamentale che emerge dal confronto tra Lewis e Pearl è quella tra *causalità di tipo* (*type causation*) e *causalità di fatto* (*token causation*). La causalità di tipo riguarda relazioni generalizzate: "il fumo causa il cancro" è un enunciato su tipi di eventi, una generalizzazione nomologica o quasi-nomologica. La causalità di fatto riguarda istanze particolari: "il fumo di Mario Rossi ha causato il suo cancro" è un enunciato su eventi particolari, su di un *token* concreto.

Lewis opera fondamentalmente a livello della causalità di fatto: la relazione causale è tra eventi particolari, localizzati spaziotemporalmente, nel mondo attuale. Il suo strumento — la dipendenza controfattuale — è relativo a un mondo specifico w. Pearl opera invece prevalentemente a livello della causalità di tipo: il modello strutturale descrive relazioni generalizzate tra variabili, la distribuzione P(**U**) è una distribuzione sulla popolazione, e le quantità come l'ACE (*Average Causal Effect*) sono medie su una popolazione. Il caso di fatto — il controfattuale individuale — è derivabile nel quadro pearliano attraverso la procedura ABP condizionata al contesto **u** specifico dell'individuo, ma richiede di conoscere (o inferire) quel contesto, il che è spesso impossibile nella pratica.

Questa distinzione ha rilevanza diretta per le applicazioni legali e morali della causalità. La responsabilità giuridica richiede tipicamente l'accertamento di causalità di fatto: non basta che il fumo *in generale* causi il cancro; occorre mostrare che *questo* fumo ha causato *questo* cancro. Il quadro pearliano fornisce, con il NCN (*Nesso Causale Necessario*) e il NCS (*Nesso Causale Sufficiente*), probabilità di causalità di fatto, ma non causalità di fatto certa a meno di conoscere il contesto individuale. Lewis, lavorando direttamente su eventi particolari, ha un accesso diretto alla causalità di fatto, ma al prezzo di non avere un metodo computazionale per ricavarla dai dati.

V.5 — L'asimmetria temporale della causalità

Un problema che tocca entrambe le teorie, ma con esiti diversi, è la *asimmetria temporale* della causalità: le cause precedono (o sono simultanee agli) effetti, e non viceversa. Da dove proviene questa asimmetria? È una caratteristica intrinseca della causalità o deriva dall'asimmetria temporale delle leggi fisiche (o dalla struttura termodinamica del mondo)?

Lewis affronta il problema in modo esplicito nel saggio del 1979: l'asimmetria temporale della causalità è ricondotta all'asimmetria del controfattuale, e questa a sua volta all'asimmetria dell'ordinamento di somiglianza tra mondi possibili. La gerarchia di priorità nell'ordinamento di somiglianza — che privilegia la conservazione della storia passata rispetto a quella futura — produce un'asimmetria nella propagazione delle dipendenze controfattuali: se c si è verificato, modificarlo "verso il passato"

richiede un miracolo grande (alterare la storia), mentre modificarlo "verso il futuro" richiede soltanto di lasciare che le conseguenze si dispieghino diversamente. Ma questa spiegazione è stata criticata da Price e altri: l'ordinamento di somiglianza sembra costruito *post hoc* per rispecchiare l'asimmetria temporale che si vuole spiegare, piuttosto che derivarla da principi indipendenti.

Pearl non affronta direttamente il problema dell'asimmetria temporale nella sua forma filosofica. Il framework degli SCM presuppone che il DAG sia aciclico, il che codifica implicitamente un ordinamento causale che rispetta il tempo (se i nodi sono indicizzati temporalmente). Ma l'ordine causale è assunto come dato nella costruzione del DAG, non derivato. La teoria non spiega *perché* le equazioni strutturali abbiano la forma f_i(PA_i, U_i) — con le cause come argomenti e gli effetti come valori — piuttosto che l'inverso. In questo senso, la domanda sull'asimmetria temporale rimane un presupposto non giustificato anche nel quadro pearliano, sebbene meno urgente perché la teoria si propone come ingegneria dell'inferenza causale piuttosto che come metafisica del tempo.

V.6 — Modularità, invarianza e il fondamento della causalità strutturale

Una delle motivazioni più profonde del programma pearliano — e di quello parallelo di Woodward nella sua teoria degli *interventismi* — è l'idea che le equazioni strutturali abbiano una proprietà che le distingue dalle mere correlazioni statistiche: la *modularità* o *invarianza*. Un'equazione strutturale V_i := f_i(PA_i, U_i) è modulare nel senso che può essere "chirurgicamente" modificata (attraverso un intervento su V_i) senza alterare le equazioni delle altre variabili. Questa modularità è l'analogo formale dell'intuizione che le relazioni causali siano robuste rispetto a perturbazioni: sapere che il sole sorge a est e che i galli cantano all'alba è una correlazione; sapere che la luce solare stimola il canto dei galli è una relazione causale che si mantiene anche se si illumina artificialmente il pollaio a mezzanotte.

James Woodward ha sviluppato questa intuizione nella teoria della causalità come *manipolability* (*manipolabilità*): c causa e se e soltanto se esiste un intervento ideale su c che modificherebbe e in modo sistematico. Questa definizione, pur nella sua circolarità apparente (un *intervento ideale* è definito in termini causali), ha il merito di fare della manipolabilità il criterio epistemico della causalità, congiungendo il concetto causale con la pratica scientifica dell'esperimento.

La critica più acuta a questo approccio viene da coloro che sostengono che la manipolabilità non è una caratteristica intrinseca della causalità ma una caratteristica delle situazioni epistemiche degli agenti: dire che c è causa di e se c è manipolabile per produrre e presuppone che gli agenti abbiano accesso fisico a c, il che non vale per fenomeni cosmologici, eventi passati o relazioni causali troppo fondamentali per essere suscettibili di intervento umano. La causalità della deriva dei continenti sull'evoluzione delle specie non è meno reale perché nessun agente può intervenire sulla posizione dei continenti. Pearl risponde a questa critica distinguendo tra *interventi fisici* e *interventi ideali*: il do-calculus descrive interventi nel senso logico di mutilazione

del modello, non nel senso pratico di azione umana. Ma questa risposta riapre il problema dell'ontologia: se l'intervento è soltanto un'operazione logica sul modello, che cosa lo distingue da un condizionamento?

La risposta di Pearl è che l'intervento e il condizionamento differiscono nella struttura del modello: condizionare su X = x lascia intatte le equazioni di X e modifica la distribuzione di **V** restringendo lo spazio campionario; intervenire su X = x elimina l'equazione di X e modifica la distribuzione di **V** alterando il meccanismo generativo. La differenza è reale e formalmente precisa, ma richiede che il modello strutturale sia già dato — e che le sue equazioni siano già interpretate come meccanismi causali, non come mere correlazioni. L'ipotesi di modularità è dunque un presupposto del framework, non un risultato: si assume che il mondo abbia una struttura causale modulare, e si costruisce il calcolo degli interventi in base a questa assunzione.

V.7 — Problemi aperti: causalità in sistemi complessi e cicli

Un limite importante sia della teoria lewisiana che di quella pearliana standard è il trattamento della causalità in sistemi con *retroazione* (*feedback*): i DAG acicli escludono per definizione i cicli causali, ma molti sistemi fisici, biologici ed economici presentano retroazione, oscillazioni, e dipendenze cicliche. Un termostato causa il calore della stanza, e il calore della stanza causa il comportamento del termostato: questa è una retroazione causale reale, non un artificio teorico.

I modelli strutturali con cicli — talvolta chiamati *Cyclic Structural Causal Models* (CSCM) — sono stati studiati da Bongers, Mooij e collaboratori, che hanno esteso il formalismo pearliano al caso ciclico sotto condizioni di esistenza e unicità della soluzione del sistema ciclico di equazioni. La condizione cruciale è che il sistema ciclico abbia un'unica soluzione stabile per ogni valore di **U**, condizione verificabile in vari modi a seconda della struttura del sistema. Quando questa condizione è soddisfatta, gli interventi e i controfattuali si calcolano in modo analogo al caso aciclico, ma la procedura di mutilazione del grafo e la propagazione delle equazioni richiedono la soluzione di un sistema di equazioni simultanee anziché la valutazione sequenziale lungo un ordinamento topologico.

Lewis, di contro, non ha sviluppato esplicitamente una teoria per la causalità ciclica. La dipendenza controfattuale in presenza di retroazione è problematica perché il condizionale "se c non si fosse verificato, e non si sarebbe verificato" può risultare circolare quando c ed e sono reciprocamente dipendenti. La teoria della causalità come influenza offre una via parziale, ma la questione rimane aperta.

Un secondo problema riguarda la causalità in sistemi con *emergenza*: fenomeni in cui proprietà macroscopiche non sono riducibili (o sono riducibili solo in modo non banale) a proprietà microscopiche. La causalità tra variabili macroscopiche — ad esempio, la pressione di un gas e il volume del contenitore — può non essere catturata da alcun modello strutturale sui costituenti microscopici (le molecole), perché la struttura causale è genuinamente di livello superiore. Questo problema, che tocca la questione generale della *riduzione* e della *supervening causation*, sfida entrambe le teorie:

Lewis può parlare di dipendenza controfattuale tra proprietà macroscopiche, ma deve rispondere alla domanda se i mondi controfattuali rilevanti sono determinati dalla fisica microscopica o da quella macroscopica; Pearl deve specificare a quale livello di granularità le equazioni strutturali sono "fondamentali".

V.8 — Verso un'integrazione: il formalismo di Halpern e Pearl per la causalità effettiva

Un tentativo rilevante di integrare la sensibilità lewisiana per la causalità di fatto con il rigore formale pearliano è la teoria della *causalità effettiva* (*actual causation*) sviluppata da Joseph Halpern e Judea Pearl, e successivamente revisionata da Halpern da solo. La domanda è: dato che il mondo si è svolto in un certo modo (con eventi C_1, …, C_k e un effetto E), quali tra i C_i sono *effettivamente* cause di E in quel particolare caso?

La definizione di Halpern-Pearl introduce due condizioni:

Condizione AC1 (condizione di fatto): C = c è vero nel modello effettivo **M** con il contesto **u** — ovvero l'evento si è effettivamente verificato.

Condizione AC2 (condizione controfattuale modificata): esiste un insieme **W** di variabili diverse da **X** e **Y**, e valori **x'** e **w**, tali che:

(a) nel modello mutilato $\mathbf{M}\{X = x', W = w\}$, *il valore di Y cambia rispetto al suo valore effettivo — ovvero* $Y\{x', w\}(\mathbf{u}) \neq y$;

(b) nel modello $\mathbf{M}\{X = x, W = w'\}$ *per qualunque sotto-insieme* $\mathbf{W'} \subseteq \mathbf{W}$ *e valori* $\boldsymbol{w'}$ *consistenti con il contesto* $\boldsymbol{u}$ *e i valori di* $\mathbf{W}$ *in* $(\boldsymbol{u})$, *il valore di Y rimane invariato — ovvero* $Y\{x, w'\}(\mathbf{u}) = y$.

La condizione AC2 è, in essenza, una versione raffinata della dipendenza controfattuale che tiene fissi alcuni "testimoni" (**W** al valore **w**) per neutralizzare gli effetti dei percorsi alternativi — la "causa di riserva" o il percorso di preemption — rendendo visibile la dipendenza del risultato dalla causa di interesse. Questa strategia — *hold fixed some variables to reveal the causal dependence* — è l'analogo strutturale della catena di dipendenze transitiva di Lewis, espressa però in termini di interventi sul modello piuttosto che di mondi possibili più vicini.

Condizione AC3 (minimalità): **X** è minimale rispetto alle condizioni AC1 e AC2 — non esiste un sottoinsieme proprio di **X** che soddisfi le stesse condizioni.

La teoria di Halpern-Pearl è stata applicata con successo a casi paradigmatici di preemption, overdetermination e causalità in sistemi complessi, e ha stimolato un dibattito ricco nella filosofia della causalità. Le sue revisioni successive hanno affinato la definizione di AC2 per evitare controesempi problematici, ma l'architettura generale — dipendenza controfattuale modificata nel quadro strutturale — rappresenta la sintesi più avanzata disponibile tra l'approccio lewisiano e quello pearliano.

V.9 — Riflessioni conclusive: struttura, meccanismo e il residuo filosofico

La traiettoria intellettuale che si è percorsa in questa monografia muove da una domanda semantica — quali sono le condizioni di verità dei controfattuali? — e appproda a una domanda ingegneristica: come si calcolano gli effetti causali dai dati? Queste due domande sono, a dispetto delle apparenze, profondamente intrecciate.

Lewis ha mostrato che i controfattuali hanno una struttura logica precisa — quella di un condizionale variabilmente rigoroso su uno spazio di mondi possibili ordinato per somiglianza — e che questa struttura è sufficiente per fondare una teoria della causalità in grado di catturare l'asimmetria temporale, la dipendenza causale, la distinzione tra cause e meri fattori causali. Il prezzo è un'ontologia robusta dei mondi possibili e un ordinamento di somiglianza la cui specificazione richiede assunzioni metafisiche sostanziali.

Pearl ha mostrato che la struttura dei controfattuali può essere *generata* da un meccanismo — il sistema di equazioni strutturali — piuttosto che assunta come primitiva, e che questo meccanismo permette di derivare, con procedura algoritmica, tutte le quantità causali di interesse a partire da dati osservativi (quando la struttura è identificabile) o di quantificarne l'incertezza residua (quando non lo è). Il prezzo è la necessità di specificare il modello strutturale, il cui fondamento epistemico rimanda a una teoria della scoperta causale che Pearl affronta con gli algoritmi di apprendimento strutturale (PC algorithm, FCI algorithm, e varianti), ma che non è mai interamente risolvibile senza assunzioni.

Quel che rimane, dopo aver percorso entrambi i programmi nella loro profondità, è un residuo filosofico irriducibile: la causalità non è solo una struttura logica, non è solo un meccanismo computazionale, e non è nemmeno soltanto una relazione tra variabili in un modello. Essa è, in primo luogo, un modo di organizzare il mondo che corrisponde a qualcosa nella realtà — alle strutture di produzione, di generazione, di dipendenza che caratterizzano i sistemi fisici, biologici, cognitivi e sociali. Il formalismo — lewisiano o pearliano — cattura questa realtà con strumenti diversi, a diversi livelli di astrazione, con diversi margini di applicabilità.

La convergenza tra i due programmi suggerisce che vi sia qualcosa di genuinamente strutturale nella causalità che entrambi i formalismi stanno cercando, ognuno a suo modo, di articolare. La divergenza suggerisce che questo quid strutturale non è ancora stato catturato in modo pienamente soddisfacente da nessuna delle due teorie, e che il problema della causalità — la sua struttura logica, la sua base ontologica, il suo ruolo epistemico — rimane uno dei problemi filosoficamente più fecondi e intellettualmente più urgenti del pensiero contemporaneo.

La matematica dei mondi possibili e quella dei grafi causali sono, in fondo, due modi di rendere precisa la stessa intuizione elementare: che le cose del mondo non stanno semplicemente insieme, ma *dipendono* le une dalle altre in modi che la sola osservazione

non riesce a svelare, e che comprendere questa dipendenza è comprendere, almeno parzialmente, *come il mondo funziona*.

Riferimenti ai testi principali discussi

David Lewis, *Counterfactuals*, Harvard University Press, 1973. — L'opera in cui viene costruita la semantica formale dei mondi possibili per i condizionali controfattuali, con il sistema **VC** e l'analisi della causalità come dipendenza controfattuale.

David Lewis, "Causation", *Journal of Philosophy* 70 (1973), pp. 556–567. — L'articolo che traduce la semantica controfattuale in una teoria della causalità, con la definizione di dipendenza causale e la catena di dipendenze.

David Lewis, "Counterfactual Dependence and Time's Arrow", *Noûs* 13 (1979), pp. 455–476. — La gerarchia di priorità nell'ordinamento di somiglianza e la derivazione dell'asimmetria temporale della causalità.

David Lewis, "Causation as Influence", *Journal of Philosophy* 97 (2000), pp. 182–197. — La revisione matura della teoria, con la causalità come influenza multidimensionale sull'effetto.

Judea Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, prima edizione 2000, seconda edizione 2009. — Il testo fondativo del programma pearliano, con la teoria degli SCM, il do-calculus, i controfattuali strutturali e la scaletta causale.

Judea Pearl, "Causal Diagrams for Empirical Research", *Biometrika* 82 (1995), pp. 669–688. — L'articolo in cui Pearl formula il do-calculus nelle sue tre regole e dimostra la formula di aggiustamento per il backdoor.

Ilya Shpitser e Judea Pearl, "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models", *AAAI* 2006. — La dimostrazione della completezza dell'algoritmo ID e del criterio dell'hedge per la non-identificabilità.

Joseph Halpern e Judea Pearl, "Causes and Explanations: A Structural-Model Approach, Part I: Causes", *British Journal for the Philosophy of Science* 56 (2005), pp. 843–887. — La teoria della causalità effettiva (actual causation) come sintesi del formalismo strutturale e dell'intuizione lewisiana.

James Woodward, *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, 2003. — La teoria interventistica della causalità come manipolabilità, parallela a Pearl ma con accento sull'esplicazione scientifica.

Sewall Wright, "Correlation and Causation", *Journal of Agricultural Research* 20 (1921), pp. 557–585. — La fonte storica della regola del cammino e dell'analisi causale strutturale in biometria.

VI. Lewis e Stalnaker: due semantiche per un solo operatore

VI.1 — Il punto di frattura: unicità versus molteplicità

La semantica dei condizionali controfattuali non nasce con Lewis: Robert Stalnaker ne aveva proposto una versione in un articolo del 1968, "A Theory of Conditionals", apparso in *Studies in Logical Theory* e destinato a diventare, insieme al lavoro lewisiano, il testo di riferimento per decenni di discussione. Le due proposte condividono l'impianto generale — mondi possibili, relazione di accessibilità, valutazione dell'antecedente nei mondi alternativi — ma divergono su un punto tecnico che ha conseguenze logiche e filosofiche di notevole portata: la questione se al controfattuale corrisponda un *unico* mondo più vicino o un *insieme* di mondi più vicini.

Stalnaker adotta la funzione di selezione f : Prop $\times$ **W** $\to$ **W**, che manda ogni coppia (antecedente, mondo di valutazione) in un *singolo* mondo — il mondo più vicino in cui l'antecedente è vero. Le condizioni di verità del condizionale $\varphi > \psi$ (Stalnaker usa $>$ invece di $\Box\!\!\rightarrow$ per distinguere il suo operatore) in un mondo w sono:

$\varphi > \psi$ è vero in w $\Leftrightarrow$ ψ è vero in $f(\varphi, w)$

con la convenzione che se φ è impossibile (non esiste nessun mondo φ-vero), il condizionale è vacuamente vero. La funzione f soddisfa le seguenti condizioni di adeguatezza:

(i) $f(\varphi, w)$ è un mondo φ-vero, oppure $f(\varphi, w) = \lambda$ (il "mondo assurdo", se φ è impossibile);

(ii) se w è esso stesso φ-vero, allora $f(\varphi, w) = w$ (il mondo attuale è il più vicino a se stesso quando l'antecedente è già verificato — *Centering*);

(iii) se $f(\varphi, w)$ è ψ-vero e $f(\psi, w)$ è φ-vero, allora $f(\varphi, w) = f(\psi, w)$ — *Condizionale di connessione*, che garantisce la coerenza del selettore su antecedenti equipossibili.

Lewis, per contro, non postula una funzione di selezione a valore singolo. Il suo operatore $\Box\!\!\rightarrow$ quantifica sull'*insieme* dei mondi φ-veri più vicini — un insieme che può contenere più di un elemento quando non esiste un unico minimo nell'ordinamento di somiglianza, o quando si adotta il sistema debole **VW** che non impone l'assioma di Limite. La differenza è dunque questa: Stalnaker assume che per ogni antecedente non contraddittorio esista un mondo *unico* che lo realizza in modo ottimale; Lewis lascia aperta la possibilità di una pluralità di mondi ottimali, e la proposizione conseguente deve essere vera in *tutti* questi mondi.

VI.2 — La validità del Conditional Excluded Middle

La conseguenza logica più immediata e più discussa di questa divergenza riguarda lo schema noto come *Conditional Excluded Middle* (CEM):

CEM: $(\varphi > \psi) \vee (\varphi > \neg\psi)$

Nel sistema stalnackeriano, questo schema è una tautologia. Poiché $f(\varphi, w)$ è un unico mondo, esso soddisfa o ψ o $\neg\psi$ (per la legge del terzo escluso classica), e quindi uno dei due condizionali è necessariamente vero.

Nel sistema lewisiano, CEM non è in generale valido. Supponiamo che vi siano esattamente due mondi φ-veri ugualmente vicini a w: u_1 in cui ψ è vero, e u_2 in cui ψ è falso. Allora $\varphi \mathrel{\square\!\!\rightarrow} \psi$ è falso in w (perché u_2 è un contro-esempio), e $\varphi \mathrel{\square\!\!\rightarrow} \neg\psi$ è pure falso in w (perché u_1 è un contro-esempio). Né il condizionale né il suo "opposto" è vero, e CEM fallisce.

Lewis argomenta che CEM *non* dovrebbe essere una tautologia, adducendo esempi come il seguente: considera un lancio di moneta equa che non è ancora avvenuto. L'enunciato "se la moneta venisse lanciata, uscirebbe testa" sembra né vero né falso — e altrettanto l'enunciato "se la moneta venisse lanciata, non uscirebbe testa". Nessuno dei due dovrebbe essere affermato come vera. Se CEM valesse, uno dei due dovrebbe essere necessariamente vero, il che sembra intuitivamente sbagliato nel caso di un evento genuinamente indeterminato.

Stalnaker risponde che CEM può essere salvato anche per eventi indeterministici, purché si adotti una interpretazione epistemica della funzione di selezione: $f(\varphi, w)$ seleziona il mondo più vicino secondo la *nostra* comprensione della situazione, non secondo una struttura metafisica oggettiva. L'indeterminismo fisico non impedisce che vi sia un mondo selezionato — semplicemente, la funzione di selezione è parzialmente indeterminata, e i nostri enunciati controfattuali sono corrispondentemente indeterminati nel valore di verità.

Questa controversia è rilevante per la teoria causale in modo non ovvio. Se CEM vale, allora per ogni coppia di eventi c ed e, o "se c non si fosse verificato, e non si sarebbe verificato" è vero, oppure "se c non si fosse verificato, e si sarebbe comunque verificato" è vero: la dipendenza o l'indipendenza causale è sempre determinata. Se CEM non vale, vi possono essere coppie di eventi per cui il rapporto causale è genuinamente indeterminato — una conseguenza che alcuni filosofi trovano gradita (l'indeterminismo causale diventa una caratteristica del mondo, non solo dell'ignoranza epistemica) e altri trovano inaccettabile (la causalità dovrebbe essere un fatto del mondo, non una questione di grado o di indeterminatezza).

VI.3 — La logica di Stalnaker: il sistema C2

Il sistema logico che Stalnaker sviluppa, noto come **C2** nella classificazione di Chellas e McKinsey, è più forte di **VC** di Lewis — in quanto comprende CEM come

tautologia — ma più debole di alcuni sistemi modali standard. I suoi assiomi caratteristici, oltre alle tautologie della logica proposizionale classica, comprendono:

RCEA (Rimpiazzamento di Antecedenti Logicamente Equivalenti): se $\vDash \varphi \leftrightarrow \chi$, allora $(\varphi > \psi) \leftrightarrow (\chi > \psi)$. Due antecedenti logicamente equivalenti selezionano lo stesso mondo, dunque il condizionale ha lo stesso valore di verità.

RCEC (Rimpiazzamento di Conseguenti Logicamente Equivalenti): se $\vDash \psi \leftrightarrow \chi$, allora $(\varphi > \psi) \leftrightarrow (\varphi > \chi)$. Segue direttamente dalla valutazione puntuale nel mondo selezionato.

CSO (Scambio Condizionale): $[(\varphi > \psi) \wedge (\psi > \varphi)] \rightarrow [(\varphi > \chi) \leftrightarrow (\psi > \chi)]$. Se φ e ψ si implicano controfattualmente a vicenda, selezionano lo stesso mondo, e quindi sono interscambiabili come antecedenti. Questo è il riflesso assiomatico della condizione (iii) sulla funzione di selezione.

CEM: $(\varphi > \psi) \vee (\varphi > \neg\psi)$. Come discusso, tautologia in Stalnaker ma non in Lewis.

CV (Conseguenza della Variazione Conditionale): $[(\varphi > \psi) \wedge \neg(\varphi > \neg\chi)] \rightarrow [(\varphi \wedge \chi) > \psi]$. Se ψ vale nel mondo più vicino a φ, e χ vale in quel mondo (poiché $\neg(\varphi > \neg\chi)$ significa che il mondo selezionato non soddisfa $\neg\chi$, ovvero soddisfa χ), allora ψ vale anche nel mondo più vicino a $\varphi \wedge \chi$ — che, per la proprietà di unicità, è lo stesso mondo.

La struttura di **C2** ha una caratterizzazione semantica elegante: è completo rispetto alla classe dei modelli di Stalnaker con funzione di selezione soddisfacente le condizioni (i)–(iii). Lewis dimostra che **C2** è un'estensione propria di **VC**: ogni tautologia di **VC** è una tautologia di **C2**, ma non viceversa (CEM è l'esempio paradigmatico).

VI.4 — La disputa sull'unicità: argomenti ontologici

La questione se esista sempre un *unico* mondo più vicino per ogni antecedente non è soltanto una scelta tecnica: ha implicazioni ontologiche che i due autori esplicitano con chiarezza.

Stalnaker è disposto ad accettare l'unicità perché ritiene che la funzione di selezione non debba essere necessariamente una relazione oggettiva tra mondi, ma possa essere parzialmente determinata dal contesto, dal contenuto dell'antecedente, e dagli interessi della conversazione. In altri termini, Stalnaker adotta una prospettiva *pragmatica* sul condizionale controfattuale: il suo valore di verità dipende non solo dalla struttura del mondo ma anche dal contesto in cui viene proferito. Questa mossa permette di mantenere CEM — perché in ogni contesto determinato vi è una selezione unica — senza impegnarsi nell'esistenza di un ordinamento di somiglianza metafisicamente oggettivo.

Lewis rifiuta questa mossa. Per lui, la funzione di selezione deve essere determinata dalla struttura del mondo, non dai contesti di proferimento; altrimenti si confonde la semantica con la pragmatica, e la verità dei controfattuali non è una questione di fatto

ma di convenzione. L'indeterminatezza nel valore di verità dei controfattuali — il fallimento di CEM — è per Lewis una caratteristica autentica di molti controfattuali, non un difetto dell'analisi. I controfattuali sono, in molti casi, genuinamente privi di valore di verità determinato, e la teoria non deve nascondere questo fatto sotto l'assunzione artificiosa dell'unicità.

Questa disputa risuona, in modo interessante, con il dibattito sulla realtà dei mondi possibili. Lewis è il paladino del *realismo modale* (*modal realism*): i mondi possibili sono entità concrete, del tutto analoghe al mondo attuale ma spaziotemporalmente separati da esso. In questa prospettiva, l'ordinamento di somiglianza ha una base ontologica robusta: i mondi sono oggetti reali, e la loro maggiore o minore somiglianza al mondo attuale è una questione di fatto. Stalnaker è invece un fautore del *concettualismo* o dell'*attualismo*: i mondi possibili sono rappresentazioni, insiemi massimi di proposizioni consistenti, o stati di cose possibili — entità astratte, non concrete. Questa ontologia più frugale è compatibile con il contestualismo semantico che Stalnaker adotta per salvare CEM.

VI.5 — Implicazioni per la teoria causale: determinismo e indeterminismo

Le conseguenze della disputa Lewis-Stalnaker per la teoria causale si manifestano con particolare nitidezza in contesti indeterministici. Nella fisica quantistica, molti eventi non hanno cause sufficienti nel senso classico: la disintegrazione di un atomo radioattivo in un istante determinato non è causata da nessuna condizione iniziale che la renda necessaria — la probabilità di disintegrazione è intrinseca e irriducibile. Come si comportano le teorie controfattuali della causalità in questo contesto?

Nel sistema stalnackeriano, poiché CEM vale, per ogni coppia di eventi c ed e o la dipendenza controfattuale è vera o la sua negazione è vera. Nel caso della disintegrazione radioattiva, questo sembrerebbe implicare che o "se non fosse stato presente questo atomo, la disintegrazione non si sarebbe verificata" è vero, oppure "se non fosse stato presente questo atomo, la disintegrazione si sarebbe verificata lo stesso" è vero — il che sembra fisicamente assurdo per eventi intrinsecamente probabilistici.

Nel sistema lewisiano, poiché CEM non vale, è possibile che nessuno dei due controfattuali sia vero: la relazione causale tra la presenza dell'atomo e la disintegrazione non è né affermata né negata dal condizionale controfattuale, che risulta semplicemente privo di valore di verità determinato. Questo esito è più confortevole dal punto di vista fisico, ma lascia aperta la domanda su come si debba intendere la causalità probabilistica.

Lewis sviluppa, in scritti successivi, una teoria della *causalità probabilistica* che integra la dipendenza controfattuale con la dipendenza probabilistica: c causa e se la probabilità di e dato c è significativamente più alta della probabilità di e in assenza di c, calcolata nei mondi più vicini. Formalmente:

c è una causa probabilistica di e ⇔ P(O(e) | O(c)) > P(O(e) | ¬O(c))

dove le probabilità sono valutate nei mondi più vicini in cui c si verifica (per il primo membro) e non si verifica (per il secondo). Questa estensione produce una teoria della causalità in grado di trattare relazioni causali stochastiche — le quali sono pervasive in biologia, epidemiologia, economia e scienze sociali — senza impegnarsi nell'esistenza di catene deterministiche sottostanti.

VI.6 — La connessione con la semantica dinamica

Un sviluppo successivo della semantica controfattuale che merita menzione è la cosiddetta *semantica dinamica* o *semantica delle informazioni in aggiornamento* (*update semantics*), elaborata da van Benthem, Veltman e, per i controfattuali in particolare, da Moss e Sack. In questa prospettiva, il significato di un condizionale non è una funzione dalle circostanze ai valori di verità, ma una *operazione di aggiornamento* sullo stato informativo dell'agente: "se φ allora ψ" aggiorna lo stato informativo corrente assumendo φ e verificando se ψ risulta derivabile dallo stato aggiornato.

Nel contesto della semantica dinamica, la funzione di selezione di Stalnaker può essere interpretata come un'operazione di *revisione delle credenze* alla Gärdenfors: l'agente revisiona le sue credenze incorporando φ nel modo minimamente disturbante, e poi verifica ψ nello stato revisionato. Questa connessione tra la semantica dei controfattuali e la teoria della revisione delle credenze — nota come *corrispondenza di Ramsey* — ha stimolato una vasta letteratura che incrocia logica, epistemologia e teoria delle decisioni.

Per la teoria causale, questa connessione suggerisce che la dipendenza controfattuale possa essere interpretata non solo come una relazione metafisica tra eventi nel mondo, ma anche come una relazione epistemica tra stati informativi dell'agente — il che riconcilia, almeno parzialmente, la prospettiva realista di Lewis con quella contestualista di Stalnaker, collocandole su piani distinti: quello ontologico e quello epistemico.

VII. Il quadro dei risultati potenziali, la mediazione causale e la scoperta strutturale

---

VII.1 — Il quadro di Rubin: risultati potenziali e SUTVA

Parallelamente al programma pearliano, e largamente indipendente da esso per decenni, si è sviluppata nella statistica applicata una tradizione di formalizzazione della causalità basata sui *risultati potenziali* (*potential outcomes*), la cui origine risale a Jerzy Neyman (1923) per i disegni sperimentali agricoli e la cui versione moderna è opera di

Donald Rubin, in una serie di contributi degli anni Settanta e Ottanta. Questo quadro — noto come modello di Neyman-Rubin o *Potential Outcomes Framework* (POF) — è oggi la lingua franca dell'econometria dell'identificazione, dell'epidemiologia causale e della valutazione delle politiche pubbliche.

Il punto di partenza è una scelta ontologica diversa da quella di Pearl: invece di costruire il formalismo attorno a un modello del meccanismo generativo (le equazioni strutturali), Rubin costruisce attorno alla nozione di *risultato potenziale*: per ogni unità i e per ogni valore t del trattamento T, esiste un risultato potenziale $Y_i(t)$ — il valore che l'esito Y assumerebbe per l'unità i se il trattamento fosse t, indipendentemente dal trattamento che i ha effettivamente ricevuto. Il risultato osservato per l'unità i che ha ricevuto il trattamento $T_i = t_i$ è:

$$Y_i^{obs} = Y_i(T_i) = Y_i(t_i)$$

mentre tutti gli altri risultati potenziali $Y_i(t)$ per $t \neq t_i$ sono controfattuali — non osservati per quella unità, per quella realizzazione dell'esperimento. Questa è la *fondamental problem of causal inference*, secondo la denominazione di Holland (1986): non si possono osservare simultaneamente più di un risultato potenziale per la stessa unità, il che rende la causalità individuale intrinsecamente non identificabile dalla sola osservazione.

L'effetto causale individuale del trattamento $T = 1$ rispetto a $T = 0$ sull'unità i è definito come:

$$\tau_i = Y_i(1) - Y_i(0)$$

Ma poiché soltanto uno dei due termini è osservabile, τ_i non è identificabile a livello individuale. Ciò che è identificabile — sotto opportune condizioni — è l'*effetto causale medio* (*Average Treatment Effect*, ATE):

$$ATE = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$$

e, in alcune circostanze, l'*effetto causale medio sui trattati* (*Average Treatment Effect on the Treated*, ATT):

$$ATT = E[Y(1) - Y(0) \mid T = 1] = E[Y(1) \mid T = 1] - E[Y(0) \mid T = 1]$$

dove il secondo termine — $E[Y(0) \mid T = 1]$, ovvero il risultato medio controfattuale che i trattati avrebbero ottenuto in assenza di trattamento — è non direttamente osservabile e costituisce la *missing data* fondamentale del problema causale.

Il presupposto che permette di collegare i risultati potenziali con la distribuzione osservata è la *Stable Unit Treatment Value Assumption* (SUTVA), articolata da Rubin in due componenti:

Non-interferenza (*no interference*): il risultato potenziale di un'unità i dipende soltanto dal trattamento ricevuto da i stessa, non dal trattamento ricevuto da altre unità. Formalmente: $Y_i(T_1, T_2, \ldots, T_n) = Y_i(T_i)$ per ogni vettore di trattamenti.

Questa condizione esclude le esternalità di rete, gli effetti spillover e le interazioni tra unità.

Versione unica del trattamento (*consistency*): esiste un'unica versione di ogni livello del trattamento. Se T_i = t, allora Y_i^{obs} = Y_i(t) senza ambiguità — non vi sono "versioni diverse" di t che potrebbero produrre risultati differenti.

VII.2 — Ignorabilità e identificazione nel POF

La condizione chiave per l'identificazione dell'ATE da dati osservativi nel POF è la *ignorabilità forte* (*strong ignorability*), introdotta da Rosenbaum e Rubin (1983):

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$$

dove $\perp\!\!\!\perp$ denota l'indipendenza statistica e X è un vettore di covariate osservate. Questa condizione — talvolta denominata *unconfoundedness* o *selection on observables* — afferma che, condizionatamente alle covariate X, l'assegnazione al trattamento è indipendente dai risultati potenziali: non vi sono confondenti non osservati che influenzino sia T che Y.

Sotto SUTVA e ignorabilità forte, l'ATE è identificato dalla distribuzione osservazionale:

$$\text{ATE} = E_X[E[Y \mid T = 1, X] - E[Y \mid T = 0, X]]$$

ovvero dalla differenza delle medie condizionate, mediata sulla distribuzione delle covariate X. Questa formula è strutturalmente identica alla formula di aggiustamento per il backdoor di Pearl, e la corrispondenza non è accidentale: la condizione di ignorabilità forte è, nel linguaggio dei DAG, esattamente la condizione che l'insieme X soddisfa il criterio del backdoor rispetto alla coppia (T, Y).

La connessione tra il POF e il formalismo pearliano è dunque profonda e quasi-biunivoca per il caso di un singolo trattamento binario senza mediatori. Pearl ha dimostrato formalmente (in collaborazione con Ilya Shpitser) che ogni problema di identificazione esprimibile nel POF è esprimibile nel do-calculus e viceversa, almeno per la classe di problemi considerata da Rubin. Tuttavia, il do-calculus ha un raggio d'azione strettamente più ampio: il POF non fornisce strumenti nativi per trattare la mediazione (effetti diretti e indiretti), la trasportabilità, i sistemi con variabili latenti complesse o le query causali che implicano la struttura del meccanismo generativo — tutte situazioni in cui il formalismo strutturale di Pearl è necessario.

VII.3 — Il propensity score e i metodi di matching

Un contributo tecnico fondamentale del POF è il *propensity score*, definito da Rosenbaum e Rubin come la probabilità condizionale di ricevere il trattamento dato le covariate:

$$e(X) = P(T = 1 \mid X)$$

Il risultato chiave è che se vale l'ignorabilità forte condizionatamente a X, allora vale anche condizionatamente al solo propensity score e(X):

$(Y(0), Y(1)) \perp\!\!\!\perp T \mid e(X)$

La dimostrazione sfrutta il fatto che, per la definizione stessa del propensity score, $T \perp\!\!\!\perp X \mid e(X)$ (il trattamento è indipendente dalle covariate dato il propensity score), e la composizione delle indipendenze condizionali nel DAG sottostante. Il propensity score è dunque un *sufficient statistic* per l'aggiustamento: invece di condizionare sull'intero vettore ad alta dimensionalità X, è sufficiente condizionare su un singolo numero reale e(X). Questo risultato ha avuto un impatto enorme sulla pratica empirica: permette di ridurre drasticamente la dimensionalità del problema di aggiustamento, rendendo praticabili metodologie come il *matching* (abbinare unità trattate e non trattate con propensity score simile), la *stratificazione* (dividere le unità in strati di propensity score omogeneo) e la *ponderazione per l'inverso della probabilità di trattamento* (*Inverse Probability of Treatment Weighting*, IPTW):

$$\text{ATE_IPTW} = E\left[T \cdot Y / e(X) \right] - E\left[(1-T) \cdot Y / (1-e(X)) \right]$$

Questa stima è *non parametrica* nel senso che non assume alcuna forma funzionale per il modello di outcome $E[Y \mid T, X]$, bensì soltanto per il modello di trattamento $e(X) = P(T \mid X)$. La sua proprietà di *consistenza doppla* (*double robustness*) — che emerge nella combinazione con un modello di outcome — è un risultato di grande rilevanza pratica: l'estimatore è consistente se almeno uno dei due modelli (il modello di trattamento o il modello di outcome) è correttamente specificato.

VIII.1 — La mediazione causale: effetti diretti e indiretti naturali

Uno degli sviluppi teorici più ricchi della causalità controfattuale è l'analisi della *mediazione causale* — lo studio di come l'effetto di un trattamento X su un esito Y si decomponga in una componente che passa attraverso un mediatore M e una componente che non vi passa. Questa decomposizione ha importanza fondamentale in molti contesti: in epidemiologia si vuole capire se l'effetto di un farmaco sul recupero passa attraverso la riduzione dell'infiammazione o attraverso meccanismi diversi; in psicologia si studia se l'effetto di un intervento sul benessere è mediato da cambiamenti nell'autostima; in economia si indaga se l'effetto dell'istruzione sui salari è mediato dalla produttività o dal segnale sociale.

La decomposizione classica dell'analisi della mediazione — dovuta a Baron e Kenny (1986) in ambito psicologico — si basa su regressioni lineari e non è fondata su un quadro causale esplicito. La versione causalmente fondata, sviluppata da Robins e Greenland (1992) e poi sistematizzata da Pearl (2001), introduce i concetti di *Effetto Diretto Naturale* (*Natural Direct Effect*, NDE) e *Effetto Indiretto Naturale* (*Natural Indirect*

Effect, NIE), che insieme sono gli strumenti standard dell'analisi causale della mediazione.

Sia X il trattamento (con valori x e x'), sia M il mediatore, sia Y l'esito. I risultati potenziali rilevanti sono:

Y_{x, M_x'} — il valore di Y quando X è impostato a x e M è impostato al valore che assumerebbe naturalmente se X fosse x'. Questa è una quantità *annidiata* o *cross-world*, poiché combina il valore di X da un "mondo" (X = x) con il valore di M da un altro "mondo" (X = x').

L'Effetto Diretto Naturale è definito come:

NDE(x, x') = E[Y_{x, M_{x'}} − Y_{x', M_{x'}}]

Questa quantità misura quanto varia Y quando X cambia da x' a x, tenendo M bloccato al valore che assumerebbe naturalmente sotto X = x'. Si tratta dell'effetto di X su Y che *non* passa attraverso M — l'effetto diretto, nel senso che prescinde dal meccanismo di mediazione.

L'Effetto Indiretto Naturale è definito come:

NIE(x, x') = E[Y_{x, M_x} − Y_{x, M_{x'}}]

Questa quantità misura quanto varia Y quando M varia dal valore naturale sotto X = x' al valore naturale sotto X = x, tenendo X fissato a x. Si tratta dell'effetto di X su Y che passa interamente attraverso la mediazione di M.

La decomposizione fondamentale della mediazione afferma che l'effetto totale di X su Y è la somma dell'effetto diretto naturale e dell'effetto indiretto naturale:

TE(x, x') = E[Y_x − Y_{x'}] = NDE(x, x') + NIE(x', x)

dove si noti la permutazione degli argomenti nell'NIE: questo riflette il fatto che la decomposizione non è perfettamente simmetrica, e che NDE(x, x') + NIE(x', x) = TE(x, x') ma NDE(x', x) + NIE(x, x') = −TE(x, x'). La decomposizione è dunque dipendente dal valore di riferimento scelto per il trattamento.

VIII.2 — Identificabilità degli effetti naturali

La caratteristica più delicata degli effetti diretti e indiretti naturali è che essi sono *quantità cross-world*: implicano il combinare risultati potenziali da configurazioni interventistiche diverse nello stesso individuo. Questo li rende potenzialmente non identificabili anche in presenza di randomizzazione del trattamento X, a differenza degli effetti causali totali.

Le condizioni di identificabilità degli effetti naturali, derivate da Pearl (2001) e Imai-Keele-Tingley (2010), richiedono:

(A1) Nessun confondente non misurato tra X e Y — ovvero l'ignorabilità forte di X rispetto a Y, condizionatamente alle covariate C: $(Y_{x,m}) \perp\!\!\!\perp X \mid C$.

(A2) Nessun confondente non misurato tra M e Y — ovvero l'ignorabilità forte di M rispetto a Y, condizionatamente a X e C: $Y_{x,m} \perp\!\!\!\perp M \mid X, C$.

(A3) Nessun confondente non misurato tra X e M — ovvero: $M_x \perp\!\!\!\perp X \mid C$.

(A4) Nessun confondente tra M e Y che sia esso stesso influenzato da X — ovvero non vi devono esistere variabili W tali che $X \to W \to Y$ e W sia confondente di M-Y. Questa condizione esclude i confondenti mediatori (*treatment-induced confounders*), la cui presenza invalida la decomposizione naturale anche quando A1–A3 sono soddisfatte.

Quando A1–A4 sono verificate, gli effetti naturali sono identificati dalla distribuzione osservazionale. La formula di identificazione per l'NDE è:

$$NDE(x, x') = \Sigma_m \Sigma_c [E[Y \mid X = x, M = m, C = c] - E[Y \mid X = x', M = m, C = c]] \cdot P(M = m \mid X = x', C = c) \cdot P(C = c)$$

e per il NIE:

$$NIE(x', x) = \Sigma_m \Sigma_c E[Y \mid X = x, M = m, C = c] \cdot [P(M = m \mid X = x, C = c) - P(M = m \mid X = x', C = c)] \cdot P(C = c)$$

Queste formule mostrano come l'effetto diretto naturale sia calcolato pesando le differenze di outcome per la distribuzione del mediatore sotto il trattamento di riferimento x', mentre l'effetto indiretto naturale sia calcolato pesando gli outcome per la differenza nelle distribuzioni del mediatore tra x e x'. La simmetria parziale tra le due formule rispecchia la struttura della decomposizione causale.

La violazione della condizione A4 — la presenza di un confondente mediatore indotto dal trattamento — è il caso più problematico in pratica. In quel caso, né l'approccio di Baron-Kenny né la formula sopra si applicano. La soluzione richiede il ricorso agli *effetti di cammino* generalizzati (*path-specific effects*) di Avin, Shpitser e Pearl (2005), che estendono la decomposizione a strutture causali con confondenti mediatori identificando esplicitamente il contributo di ogni cammino causale nel DAG.

IX.1 — La scoperta causale: il problema inverso

I problemi affrontati finora — identificazione degli effetti causali, calcolo degli interventi, derivazione dei controfattuali — sono problemi *diretti*: data una struttura causale (un DAG o un modello strutturale), si vuole calcolare una quantità causale di interesse. Il problema *inverso* è altrettanto fondamentale, sia teoricamente che praticamente: data una distribuzione osservazionale $P(\mathbf{V})$, è possibile inferire la struttura causale sottostante — il DAG che l'ha generata?

Questo è il problema della *scoperta causale* (*causal discovery*), affrontato sistematicamente da Peter Spirtes, Clark Glymour e Richard Scheines nella monografia *Causation, Prediction, and Search* (prima edizione 1993, seconda edizione 2000), che costituisce, insieme ai lavori di Pearl, l'altro pilastro della teoria causale computazionale contemporanea.

Il punto di partenza è l'osservazione che, in generale, la distribuzione P(**V**) non identifica univocamente il DAG che l'ha generata. Più DAG possono generare la stessa distribuzione osservazionale — essi sono detti *Markov equivalenti* — e costituiscono insieme una *classe di equivalenza di Markov*. Due DAG **G**_1 e **G**_2 sono Markov equivalenti se e soltanto se implicano le stesse indipendenze condizionali, ovvero se la loro d-separazione produce le stesse asserzioni di indipendenza. La caratterizzazione grafica della Markov equivalenza è data dal teorema di Verma e Pearl: due DAG sono Markov equivalenti se e soltanto se hanno lo stesso scheletro (stesso insieme di archi, senza considerare l'orientamento) e gli stessi *v-strutture* (*collider non-adiacenti*, detti anche *immoralities*), ovvero le stesse configurazioni del tipo V_i → V_k ← V_j con V_i e V_j non adiacenti.

La struttura grafica che rappresenta una classe di equivalenza di Markov è il *CPDAG* (*Completed Partially Directed Acyclic Graph*), in cui alcuni archi sono orientati — quelli il cui orientamento è invariante in tutta la classe — e altri sono non orientati — quelli il cui orientamento è compatibile con più DAG nella classe.

IX.2 — L'algoritmo PC

L'algoritmo PC (dai nomi di **P**eter Spirtes e **C**lark Glymour) è la procedura di scoperta causale più classica e ampiamente studiata. Opera in due fasi principali.

Fase 1 — Scheletro: si parte dal grafo completo su **V** (tutti i nodi collegati a tutti gli altri) e si rimuovono iterativamente gli archi per cui si trova un insieme di separazione — un insieme $S_{ij} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$ tale che $V_i \perp\!\!\!\perp_P V_j \mid S_{ij}$ sia verificata (in senso statistico, tipicamente attraverso un test di indipendenza condizionale). Il risultato è il *grafo non orientato* che rispecchia le adiacenze causali del DAG sottostante.

La complessità computazionale della Fase 1 dipende crucialmente dall'ordine in cui si testano le indipendenze: l'algoritmo PC testa prima le indipendenze marginali ($|S| = 0$), poi le indipendenze condizionate su una variabile ($|S| = 1$), poi su due ($|S| = 2$), e così via, procedendo in ordine crescente di cardinalità degli insiemi di separazione. Questa strategia — lo *skeleton algorithm* — riduce notevolmente il numero di test necessari sotto l'assunzione di *sparsità* del grafo (pochi archi rispetto al numero di nodi), che è verificata in molte applicazioni pratiche.

Fase 2 — Orientamento: una volta determinato lo scheletro, si orientano gli archi seguendo le regole di Meek (1995), che garantiscono che il risultato sia il CPDAG della classe di equivalenza di Markov. Le regole principali sono:

Regola v-struttura: per ogni tripla V_i — V_k — V_j con V_i e V_j non adiacenti, se S_{ij} non contiene V_k, allora V_k è un collider per quella tripla, e si orienta V_i → V_k ← V_j.

Regola R1 (non-collider): se V_i → V_k — V_j e V_i e V_j non sono adiacenti, allora l'arco V_k — V_j deve essere orientato V_k → V_j (altrimenti si creerebbe una nuova v-struttura non supportata dai dati).

Regola R2 (aciclichità): se esiste un cammino orientato V_i → … → V_j e un arco V_i — V_j, allora l'arco deve essere orientato V_i → V_j per preservare l'aciclichità.

Regola R3 e R4: regole aggiuntive per casi in cui R1 e R2 non sono sufficienti a orientare univocamente tutti gli archi compatibili con le v-strutture.

La correttezza dell'algoritmo PC — il fatto che produce il CPDAG corretto sotto l'ipotesi di fedeltà e un numero sufficiente di dati — è dimostrata da Spirtes-Glymour-Scheines e da Meek. La consistenza — il fatto che la probabilità di produrre il CPDAG errato tende a zero con l'aumentare della dimensione campionaria — è dimostrata sotto ipotesi di regolarità sui test di indipendenza condizionale.

IX.3 — L'algoritmo FCI e i confondenti latenti

L'algoritmo PC presuppone l'assenza di variabili latenti: il DAG sottostante ha soltanto variabili osservate. Quando questa assunzione è violata — e nella pratica lo è quasi sempre — occorre un algoritmo diverso, in grado di produrre una rappresentazione della struttura causale che includa i confondenti latenti.

L'algoritmo FCI (*Fast Causal Inference*), sviluppato da Spirtes, Glymour e Scheines, estende PC al caso con variabili latenti producendo un *PAG* (*Partial Ancestral Graph*), un tipo di grafo che usa archi con marcatori di coda distinti (punta di freccia →, coda piena —, cerchio ○) per rappresentare l'incertezza sull'orientamento e sulla presenza di confondenti latenti. Un arco V_i ○→ V_j indica che V_i è un possibile genitore o un possibile confondente di V_j; un arco V_i ↔ V_j indica la presenza certa di un confondente latente tra V_i e V_j.

Il PAG rappresenta una classe di equivalenza di *MAG* (*Maximal Ancestral Graphs*): grafi che codificano tutte le strutture causali (con variabili latenti marginalizzate) compatibili con le indipendenze condizionali osservate. La relazione tra PAG, MAG e DAG semi-markoviani è la seguente: ogni DAG semi-markoviano è rappresentabile come un MAG mediante l'eliminazione delle variabili latenti e la sostituzione delle loro influenze con archi bidirezionali; ogni PAG rappresenta la classe di equivalenza di MAG che condividono scheletro e v-strutture nel grafo degli antenati.

La complessità teorica del FCI e la difficoltà pratica di stimare correttamente le indipendenze condizionali in presenza di confondenti latenti hanno stimolato lo sviluppo di varianti e miglioramenti, tra cui RFCI (*Really Fast Causal Inference*) per insiemi di dati di grandi dimensioni, e metodi ibridi che combinano la ricerca basata su

indipendenze condizionali con la ricerca basata su massimizzazione dello *score* (come l'algoritmo GFCI).

IX.4 — Identificabilità della classe di equivalenza e informazioni aggiuntive

In generale, i metodi basati sull'analisi delle indipendenze condizionali possono identificare soltanto la classe di equivalenza di Markov — non il DAG specifico — a meno di disporre di informazioni aggiuntive che rompano la simmetria all'interno della classe. Le principali fonti di tali informazioni sono:

Ordine temporale: se si sa che V_i precede temporalmente V_j, allora l'arco tra loro non può essere orientato V_i ← V_j. L'ordine temporale tra variabili, quando disponibile, riduce drasticamente la classe di equivalenza e spesso identifica univocamente il DAG.

Asimmetria della distribuzione degli errori: nei modelli lineari non gaussiani (*Linear Non-Gaussian Acyclic Models*, LiNGAM), l'orientamento causale è identificabile dalla sola distribuzione osservazionale senza test di indipendenza condizionale aggiuntivi. Il risultato chiave, dimostrato da Shimizu e collaboratori (2006), è che se il modello è lineare e gli errori sono non gaussiani, la direzione causale è identificabile dal segno dell'asimmetria (skewness) delle distribuzioni dei residui: il DAG corretto è quello per cui i residui di ogni variabile regressata sulle sue candidate cause sono indipendenti da quelle cause.

Formalmente, nel modello lineare $V = B \cdot V + \varepsilon$ con B triangolare inferiore, se gli ε_i sono non gaussiani e mutualmente indipendenti, allora la distribuzione di $\mathbf{V}$ identifica univocamente B (e dunque il DAG) fino a permutazione delle variabili. Il metodo ICA (*Independent Component Analysis*) applicato ai dati osservati produce una stima di B dalla quale si ricava la struttura causale.

Non-linearità delle equazioni strutturali: nei modelli causali additivi non lineari (*Additive Noise Models*, ANM), anche con errori gaussiani, l'orientamento causale è identificabile se le relazioni funzionali sono non lineari. Il criterio è analogo a quello di LiNGAM: il DAG corretto è quello per cui i residui di ogni variabile rispetto alla funzione strutturale (stimata non parametricamente) sono indipendenti dagli input.

IX.5 — Invarianza causale e generalizzazione tra ambienti

Un approccio recente alla scoperta causale, che collega direttamente la struttura causale con la nozione di *invarianza* discussa in §V.6, è il metodo delle *Invariant Causal Prediction* (ICP), proposto da Peters, Mooij, Bühlmann e Schölkopf (2016). L'idea fondante è che le equazioni strutturali — a differenza delle mere regressioni — siano invarianti attraverso diversi ambienti (diversi interventi, diverse popolazioni, diversi contesti): il meccanismo causale $V_i := f_i(PA_i, U_i)$ rimane lo stesso anche quando la distribuzione di PA_i cambia a causa di interventi esterni.

Formalmente, si dispone di dati provenienti da k ambienti E_1, …, E_k, ciascuno corrispondente a un diverso intervento o condizione di raccolta. La distribuzione di Y condizionatamente ai suoi genitori causali diretti PA_Y è invariante attraverso tutti gli ambienti:

P^{E_j}(Y | PA_Y) = P^{E_l}(Y | PA_Y) per ogni j, l ∈ {1,…,k}

mentre la distribuzione di Y condizionatamente a variabili che non includono tutti i suoi genitori causali *non* è in generale invariante. Il metodo ICP procede testando, per ogni sottoinsieme $S \subseteq \mathbf{V} \setminus \{Y\}$, se la distribuzione condizionale $P(Y \mid S)$ è invariante attraverso gli ambienti; l'intersezione di tutti i sottoinsiemi S che passano il test di invarianza è il miglior candidato per PA_Y.

Il legame con il programma pearliano è diretto: l'invarianza della distribuzione condizionale di Y dato i suoi genitori è esattamente la proprietà che definisce le equazioni strutturali come *meccanismi causali* piuttosto che come *correlazioni statistiche*. ICP traduce questa proprietà in un criterio empirico testabile, aprendo la via all'identificazione della struttura causale da dati provenienti da contesti multipli — una situazione molto più comune, nella pratica scientifica, di quella in cui si dispone di dati da un unico contesto omogeneo.

Appendice A — Logiche modali, sistemi di Kripke e la collocazione della logica controfattuale

A.1 — La gerarchia delle logiche modali normali

Per comprendere con precisione la posizione della logica controfattuale di Lewis nel panorama delle logiche modali, è necessario richiamare la struttura della gerarchia dei sistemi modali normali, partendo dal sistema base **K** e risalendo verso sistemi via via più forti.

Una *logica modale normale* è un'estensione della logica proposizionale classica che aggiunge un operatore di necessità $\Box$ (e il suo duale $\Diamond = \neg\Box\neg$, la possibilità), soggetto ai seguenti requisiti minimi:

Assioma K (distribuzione): $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$. Questo afferma che se è necessario che φ implichi ψ, e φ è necessario, allora ψ è necessario. È l'assioma che caratterizza tutte le logiche modali normali.

Regola di Necessitazione: se $\vdash \varphi$, allora $\vdash \Box\varphi$. Le tautologie sono necessariamente vere.

Oltre a **K**, i sistemi principali si caratterizzano per l'aggiunta di uno o più assiomi:

T (riflessività): $\Box\varphi \rightarrow \varphi$. Ciò che è necessario è vero. Corrisponde alla condizione che la relazione di accessibilità tra mondi sia riflessiva: $\forall w \in \mathbf{W} : wRw$.

B (simmetria): $\varphi \rightarrow \Box\Diamond\varphi$. Corrisponde alla condizione che la relazione di accessibilità sia simmetrica: $wRv \Rightarrow vRw$.

4 (transitività): $\Box\varphi \rightarrow \Box\Box\varphi$. Corrisponde alla condizione che la relazione di accessibilità sia transitiva: $wRv \wedge vRu \Rightarrow wRu$.

5 (euclideanità): $\Diamond\varphi \rightarrow \Box\Diamond\varphi$. Corrisponde alla condizione che la relazione di accessibilità sia euclidea: $wRv \wedge wRu \Rightarrow vRu$.

I sistemi ottenuti per combinazione di questi assiomi con **K** sono: **KT** (detto anche **T**), **KT4** (detto **S4**), **KT4B** = **KT5** (detto **S5**). **S5** è il sistema per eccellenza della necessità logica o metafisica; **S4** modella la necessità epistemica o la dimostrabilità; **T** è il sistema minimale adatto alla necessità nomologica.

La semantica di Kripke per questi sistemi è la seguente. Un *frame di Kripke* è una coppia $\langle \mathbf{W}, R \rangle$ dove $\mathbf{W}$ è un insieme non vuoto di mondi e $R \subseteq \mathbf{W} \times \mathbf{W}$ è la relazione di accessibilità. Un *modello di Kripke* è una tripla $\langle \mathbf{W}, R, \mathbf{V} \rangle$ con $\mathbf{V} : \text{Prop} \times \mathbf{W} \rightarrow \{0,1\}$ una funzione di valutazione. La verità di $\Box\varphi$ in w è:

$\Box\varphi$ è vero in w $\Leftrightarrow \forall v \in \mathbf{W} : wRv \Rightarrow \varphi$ è vero in v

Il teorema fondamentale di completezza di Kripke stabilisce che ogni sistema modale normale è completo rispetto alla classe dei frame che soddisfano le condizioni relazionali corrispondenti: **T** rispetto ai frame riflessivi, **S4** rispetto ai frame riflessivi e transitivi (frame di preordine), **S5** rispetto ai frame di equivalenza.

A.2 — Le logiche condizionali come logiche modali non-normali

La logica controfattuale di Lewis appartiene alla famiglia delle *logiche condizionali* o *logiche modali non-normali*. L'operatore $\Box\!\!\rightarrow$ non è un operatore di necessità standard, perché la regola di Necessitazione non vale per esso: non è vero in generale che se ψ è una tautologia allora $\varphi \Box\!\!\rightarrow \psi$ è una tautologia. Più precisamente, $\Box\!\!\rightarrow$ è un operatore *binario*, il che lo distingue strutturalmente dall'operatore unario $\Box$ della logica modale standard.

La famiglia di logiche a cui appartiene Lewis è quella delle *logiche di selezione* o *logiche di condizionale variabilmente rigoroso*, studiate sistematicamente da Chellas nel libro *Modal Logic: An Introduction* (1980). Un sistema di selezione è un frame $\langle \mathbf{W}, \{\mathbf{f}w\}\{w \in \mathbf{W}\} \rangle$ dove $\mathbf{f}_w : \mathscr{P}(\mathbf{W}) \rightarrow \mathscr{P}(\mathbf{W})$ è una *funzione di selezione* che manda ogni proposizione (insieme di mondi) nell'insieme dei mondi "selezionati" come rilevanti per la valutazione del condizionale. Le condizioni di verità di $\varphi \Box\!\!\rightarrow \psi$ in w sono:

$\varphi \Box\!\!\rightarrow \psi$ è vero in w $\Leftrightarrow \mathbf{f}_w(|\varphi|) \subseteq |\psi|$

dove $|\varphi| = \{v \in \mathbf{W} : \varphi \text{ è vero in } v\}$ è la *proposizione* espressa da φ. Nella formulazione di Lewis, $\mathbf{f}_w(|\varphi|)$ è l'insieme dei mondi $|\varphi|$-veri più vicini a w nell'ordinamento $\leq_w$ — il selettore che nel testo principale si è indicato con $\mathbf{f}_\varphi(w)$.

Le proprietà dell'operatore $\Box\!\rightarrow$ dipendono dalle proprietà della funzione di selezione. Lewis impone le seguenti condizioni, corrispondenti agli assiomi del sistema **VC**:

(f1) $\mathbf{f}_w(|\varphi|) \subseteq |\varphi|$ — il selettore seleziona solo mondi φ-veri (*Success*);

(f2) se $w \in |\varphi|$, allora $w \in \mathbf{f}_w(|\varphi|)$ — il mondo attuale è selezionato quando φ vi è vero (*Centering debole*);

(f3) se $w \in |\varphi|$, allora $\mathbf{f}_w(|\varphi|) = \{w\}$ — il mondo attuale è l'unico selezionato se φ vi è vero (*Centering forte*);

(f4) se $\mathbf{f}_w(|\varphi|) \cap |\psi| \neq \varnothing$, allora $\mathbf{f}_w(|\varphi| \cap |\psi|) \subseteq \mathbf{f}_w(|\varphi|)$ — *condizione di Limite inferiore*;

(f5) $\mathbf{f}_w(|\varphi| \cap |\psi|) \cap |\varphi| \subseteq \mathbf{f}_w(|\varphi|)$ — *condizione di Monotonicità limitata*.

Le condizioni (f1)–(f5) corrispondono esattamente agli assiomi di **VC**. La gerarchia dei sistemi condizionali è:

$$\mathbf{ID} \subset \mathbf{VW} \subset \mathbf{V} \subset \mathbf{VN} \subset \mathbf{VT} \subset \mathbf{VC} \subset \mathbf{C2}$$

dove ogni sistema a sinistra è esteso in modo proprio da quello a destra. **C2** di Stalnaker aggiunge la condizione di unicità del selettore: $|\mathbf{f}_w(|\varphi|)| = 1$ per ogni φ non contraddittorio, corrispondente all'assioma CEM.

A.3 — Il teorema di completezza per VC e C2

Il teorema di completezza per **VC** afferma che ogni tautologia è dimostrabile nel sistema assiomatico corrispondente, e viceversa. La dimostrazione procede costruendo, a partire da ogni insieme di formule massimale consistente Γ, un *modello canonico* in cui:

I mondi sono gli insiemi massimali consistenti di formule del linguaggio. L'ordinamento $\leq_w$ è definito in modo che $u \leq_w v$ se e soltanto se ogni formula del tipo $(\varphi \Box\!\rightarrow \psi) \in w$ che sia vera in v è anche vera in u (ovvero u è "almeno altrettanto vicino" a w quanto v, nel senso delle formule controfattuali che il modello deve soddisfare). La funzione di valutazione assegna il valore 1 alle formule atomiche contenute nel mondo corrispondente.

Per **C2**, la dimostrazione dell'ulteriore assioma CEM nella classe di modelli di Stalnaker è immediata per l'unicità del selettore: se $\mathbf{f}_w(|\varphi|) = \{u\}$ per un unico u, allora u soddisfa per la logica classica o ψ o $\neg\psi$, e dunque uno dei due condizionali $\varphi > \psi$ o $\varphi > \neg\psi$ è necessariamente vero in w.

A.4 — SCM e S5: necessità nomologica nei modelli strutturali

In un modello strutturale **M** in cui le equazioni strutturali esprimono leggi nomologiche, l'operatore di necessità nomologica □ può essere interpretato come quantificazione su tutti i mondi che condividono le stesse equazioni strutturali ma differiscono nei valori delle variabili esogene. Formalmente, il frame $\langle \mathbf{W}\boldsymbol{M}, R\mathbf{M} \rangle$ in cui $\mathbf{W}\boldsymbol{M} = dom(\boldsymbol{U})$ *e* $R\mathbf{M} = \mathbf{W}\boldsymbol{M} \times \boldsymbol{W}\mathbf{M}$ (relazione di accessibilità universale) corrisponde al sistema modale **S5** con rispetto alla distribuzione delle variabili esogene:

□φ è vero in **u** (nel modello **M**) ⇔ φ è vero per ogni $\mathbf{u}' \in \mathrm{dom}(\mathbf{U})$

Le leggi nomologiche del modello — le equazioni strutturali f_i — sono necessarie in questo senso: $\forall\, \mathbf{u} \in \mathrm{dom}(\mathbf{U})$: V_i(**u**) = f_i(PA_i(**u**), U_i(**u**)), e questo vale in *ogni* mondo del modello. Le relazioni controfattuali — derivate mediante la procedura ABP — corrispondono a quantificazione sul sottoinsieme di mondi in cui X = x (cioè sul modello mutilato **M**_x), in accordo con la semantica delle logiche condizionali su frame di selezione. La dimostrazione formale di questa corrispondenza — che i controfattuali derivabili in un SCM soddisfano gli assiomi di **VC** — costituisce il risultato principale di Briggs (2012).

Appendice B — Teoria dei grafi causali: DAG, MAG, CPDAG, d-separazione

B.1 — Definizioni fondamentali

Un *grafo orientato* è una coppia $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ dove **V** è un insieme finito non vuoto di nodi e $\mathbf{E} \subseteq (\mathbf{V} \times \mathbf{V}) \setminus \{(v,v)\}$ è l'insieme degli archi orientati privi di cappi. Per ogni arco $(V_i, V_j) \in \mathbf{E}$ — scritto $V_i \to V_j$ — diciamo che V_i è un *genitore* di V_j e V_j è un *figlio* di V_i. L'insieme dei genitori di V_i è $PA_i = \{V_j : V_j \to V_i \in \mathbf{E}\}$.

Un *cammino* in **G** da V_i a V_j è una sequenza di nodi distinti $(V_{k_0}, V_{k_1}, \ldots, V_{k_m})$ con $V_{k_0} = V_i$, $V_{k_m} = V_j$, in cui ogni coppia consecutiva $(V_{k_l}, V_{k_{l+1}})$ è connessa da un arco (in qualunque direzione). Un *cammino orientato* da V_i a V_j richiede che ogni arco sia nella direzione $V_{k_l} \to V_{k_{l+1}}$. Un *ciclo diretto* è un cammino orientato che parte e arriva allo stesso nodo. Un *grafo aciclico orientato* (DAG) è un grafo orientato privo di cicli diretti.

L'insieme degli antenati di $\mathbf{S} \subseteq \mathbf{V}$ è:

$$\mathrm{An}(\mathbf{S})_{\mathbf{G}} = \{V_i \in \mathbf{V} : \exists\ V_j \in \mathbf{S} \text{ con cammino orientato } V_i \dashrightarrow V_j \text{ in } \mathbf{G}\} \cup \mathbf{S}$$

e $\mathrm{De}(\mathbf{S})_{\mathbf{G}}$ è l'insieme dei discendenti, definito simmetricamente. Si noti che per convenzione ogni nodo è antenato e discendente di se stesso.

B.2 — Condizione di Markov e fattorizzazione causale

Una distribuzione P su **V** soddisfa la *condizione di Markov causale* per il DAG **G** se e soltanto se:

$\forall$ V_i $\in$ **V** : V_i $\perp\!\!\!\perp$ *P (Non-De(V_i)***G** \ PA_i) | PA_i

dove Non-De(V_i)$_{\boldsymbol{G}}$ = $\boldsymbol{V}$ \ *De({V_i})***G** è l'insieme dei non-discendenti di V_i.

Questa condizione è equivalente alla fattorizzazione:

P(V_1, …, V_n) = $\prod$_{i=1}^{n} P(V_i | PA_i)

Dimostrazione dell'equivalenza: sia π = (V_{π(1)}, …, V_{π(n)}) un ordinamento topologico di **G** (sempre esistente nei DAG). Per la regola della catena:

P(V_1, …, V_n) = $\prod${*i=1*}^{*n*} *P(V*{π(i)} | V_{π(1)}, …, V_{π(i-1)})

In un ordinamento topologico, {V_{π(1)}, …, V_{π(i-1)}} $\subseteq$ Non-De(V_{π(i)})$_{\boldsymbol{G}}$. *Per la condizione di Markov applicata a V*{π(i)}:

V_{π(i)} $\perp\!\!\!\perp$ *P ({V*{π(1)}, …, V_{π(i-1)}} \ PA_{π(i)}) | PA_{π(i)}

Dunque P(V_{π(i)} | V_{π(1)}, …, V_{π(i-1)}) = P(V_{π(i)} | PA_{π(i)}), e la fattorizzazione segue.

Per la direzione inversa, data la fattorizzazione, si può verificare la condizione di Markov mediante marginalizzazione: la distribuzione marginale di V_{π(i)} condizionata a PA_{π(i)} si ottiene sommando sulle variabili restanti nei fattori che non coinvolgono V_{π(i)}, e il risultato è indipendente dai non-discendenti al di là dei genitori. □

B.3 — La d-separazione: definizione formale

Sia **G** = (**V**, **E**) un DAG. Siano **X**, **Y**, **Z** $\subseteq$ **V** insiemi a due a due disgiunti. Un cammino p tra un nodo di **X** e un nodo di **Y** è *bloccato* da **Z** se contiene almeno uno dei seguenti tipi di configurazioni:

(B1) Un nodo V_k $\in$ **Z** che compare in p come *non-collider*: V_k è una catena (V_i $\rightarrow$ V_k $\rightarrow$ V_j o V_i $\leftarrow$ V_k $\leftarrow$ V_j) o una forchetta (V_i $\leftarrow$ V_k $\rightarrow$ V_j). Il condizionamento su V_k blocca il flusso associativo.

(B2) Un nodo V_k che compare in p come *collider* (V_i $\rightarrow$ V_k $\leftarrow$ V_j), tale che V_k $\notin$ **Z** e nessun discendente di V_k appartiene a **Z**. Il collider non condizionato blocca il flusso.

Gli insiemi **X** e **Y** sono *d-separati* da **Z** in **G** — scritto (**X** $\perp\!\!\!\perp$ **Y** | **Z**)_**G** — se ogni cammino tra qualsiasi V_x $\in$ **X** e qualsiasi V_y $\in$ **Y** è bloccato da **Z**.

Osservazione: il fatto che il condizionamento su un collider *apra* il flusso (annullando il blocco della condizione B2) è il fondamento formale del fenomeno di *collider bias* o *selection bias*: condizionare su una variabile che è un collider tra due cause crea un'associazione statistica tra quelle cause anche in assenza di qualunque relazione causale tra loro. Questo è il meccanismo sottostante al *paradosso di Berkson* in epidemiologia.

B.4 — Teorema di completezza della d-separazione

Teorema (Markov per DAG, direzione principale): sia P una distribuzione che si fattorizza rispetto al DAG **G**. Per ogni tripla **X**, **Y**, **Z** con (**X** ⊥⊥ **Y** | **Z**)_**G**, si ha X ⊥⊥_P Y | **Z**.

Dimostrazione mediante il criterio del grafo morale. Il *grafo morale* **G**^m di un DAG **G** è il grafo non orientato ottenuto: (a) per ogni nodo V_k con genitori PA_k = {V_{i_1}, …, V_{i_r}}, aggiungere archi tra ogni coppia (V_{i_j}, V_{i_l}) con j ≠ l (*matrimonio dei genitori*); (b) sostituire ogni arco orientato con uno non orientato.

Lemma ausiliario (Lauritzen-Dawid-Larsen-Leimer, 1990): se (**X** ⊥⊥ **Y** | **Z**)***G*** *allora* ***X*** *e* ***Y*** *sono separati da* ***Z*** *nel sottografo morale di* ***G*** *ristretto all'insieme degli antenati* $An(\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z})_{\mathbf{G}}$.

Dato il Lemma, la dimostrazione del teorema procede applicando il *lemma di Hammersley-Clifford* al grafo non orientato: in un grafo non orientato con distribuzione positiva che si fattorizza secondo le clique del grafo, la separazione di **X** da **Y** dato **Z** implica X ⊥⊥_P Y | **Z**. La positività di P (che si può assumere senza perdita di generalità con una perturbazione arbitrariamente piccola) garantisce l'applicabilità del lemma.

La fattorizzazione di P rispetto a **G** implica la fattorizzazione rispetto al grafo morale degli antenati — perché ogni clique del grafo morale corrisponde a un insieme di variabili che compaiono insieme in almeno un fattore P(V_i | PA_i) della fattorizzazione causale. La separazione grafica produce quindi l'indipendenza statistica. □

Teorema di completezza (Verma, 1993, sotto fedeltà): se P è fedele a **G**, allora (**X** ⊥⊥ **Y** | **Z**)_**G** ⇔ X ⊥⊥_P Y | **Z**.

La direzione inversa — che sotto fedeltà ogni indipendenza statistica corrisponde a una d-separazione — si dimostra per contrappositivo: se **X** e **Y** sono d-connessi dato **Z** (non d-separati), allora sotto fedeltà P(**X**, **Y** | **Z**) non è fattorizzabile come P(**X** | **Z**) · P(**Y** | **Z**), ovvero X e Y sono statisticamente dipendenti dato **Z**.

B.5 — Caratterizzazione della Markov equivalenza e costruzione del CPDAG

Teorema (Verma-Pearl, 1990): due DAG **G**_1 e **G**_2 su **V** sono Markov equivalenti se e soltanto se:

(i) hanno lo stesso *scheletro* — stesso insieme di coppie di nodi adiacenti, ignorando l'orientamento;

(ii) hanno le stesse *v-strutture* — per ogni tripla (V_i, V_k, V_j) con V_i e V_j non adiacenti in **G**_1 (e quindi in **G**_2 per (i)), V_k è collider in **G**_1 se e soltanto se è collider in **G**_2.

Dimostrazione (schema): per la direzione (⟹), si mostra che se **G**_1 e **G**_2 hanno le stesse d-separazioni, allora hanno lo stesso scheletro (perché l'adiacenza equivale alla d-connessione per ogni insieme Z) e le stesse v-strutture (perché una v-struttura produce una d-connessione che è assente se si orienta il collider come non-collider). Per la direzione (⟸), si verifica per esaustione che due DAG con lo stesso scheletro e le stesse v-strutture producono le medesime asserzioni di d-separazione, applicando le regole (B1) e (B2) a ogni possibile cammino.

La costruzione del *CPDAG* da un DAG **G** procede in due fasi:

Fase 1: si orientano le v-strutture — per ogni tripla (V_i, V_k, V_j) con V_i e V_j non adiacenti e V_k adiacente a entrambi, se V_k è collider in **G** si orienta V_i → V_k ← V_j nel CPDAG.

Fase 2: si applicano iterativamente le quattro regole di Meek (1995) fino a convergenza. Le prime due regole sono le più frequentemente attive:

Regola R1: se V_i → V_k — V_j nel CPDAG corrente e V_i e V_j non sono adiacenti, orientare V_k → V_j. Motivazione: l'alternativa V_k ← V_j creerebbe la v-struttura V_i → V_k ← V_j, che non è supportata dai dati (V_i e V_j non erano un collider nello scheletro originale).

Regola R2: se esiste il cammino V_i → V_k → V_j e l'arco non orientato V_i — V_j nel CPDAG corrente, orientare V_i → V_j. Motivazione: l'alternativa V_i ← V_j creerebbe un ciclo diretto V_i → V_k → V_j → V_i, violando l'aciclichità.

Regola R3: se V_k — V_i — V_j e V_k — V_l — V_j, V_k → V_i, V_l → V_i, V_k e V_l non adiacenti, e V_i — V_j, orientare V_i → V_j.

Regola R4: se V_k → V_l → V_j e V_k → V_i — V_j, orientare V_i → V_j.

La terminazione è garantita perché ogni applicazione delle regole orienta almeno un arco, e il numero di archi è finito. La correttezza — ovvero il fatto che il CPDAG prodotto è esattamente il rappresentante canonico della classe di equivalenza di **G** — è dimostrata da Meek per induzione sulla struttura delle regole.

B.6 — MAG, PAG e d-separazione con variabili latenti

Un *MAG* (**M**aximal **A**ncestral **G**raph) su un insieme di variabili osservate **O** è un grafo misto con archi orientati (→) e bidirezionali (↔) tale che:

(M1) Non vi sono cicli diretti.

(M2) Per ogni arco bidirezionale $V_i \leftrightarrow V_j$, $V_i \notin An(V_j)$ e $V_j \notin An(V_i)$ nel MAG — le variabili collegate da un arco bidirezionale non sono in relazione antenato-discendente.

(M3) Per ogni coppia (V_i, V_j) non adiacente nel MAG, esiste un insieme $\mathbf{Z} \subseteq \mathbf{O}$ tale che V_i e V_j sono d-separati da $\mathbf{Z}$ nel MAG — *condizione di massimalità*.

La corrispondenza tra DAG semi-markoviani (DAG con variabili latenti $\mathbf{L}$) e MAG è stabilita dalla proiezione: dato un DAG semi-markoviano $\langle \mathbf{O} \cup \mathbf{L}, \mathbf{E} \rangle$, il MAG su $\mathbf{O}$ ha:

un arco orientato $V_i \to V_j$ per ogni $V_i \in \mathbf{O}$ tale che V_i è un antenato di V_j in $\mathbf{O}$ nel DAG originale, con $V_i \in PA_j$ oppure con tutti i cammini orientati da V_i a V_j che passano soltanto per variabili latenti in $\mathbf{L}$;

un arco bidirezionale $V_i \leftrightarrow V_j$ per ogni coppia $(V_i, V_j) \in \mathbf{O}$ tali che esiste una variabile latente $L \in \mathbf{L}$ che è antenato di entrambi senza che né V_i né V_j sia antenato dell'altro.

La d-separazione si estende ai MAG attraverso le medesime regole (B1) e (B2), trattando i nodi degli archi bidirezionali come collider nei percorsi. Il *teorema di completezza per MAG* (Richardson-Spirtes, 2002) afferma che, sotto la condizione di fedeltà estesa, le indipendenze condizionali di P sull'insieme $\mathbf{O}$ coincidono esattamente con le asserzioni di d-separazione nel MAG corrispondente.

Il *PAG* rappresenta la classe di equivalenza di MAG mediante archi con estremità marcate in quattro modi: punta di freccia ($>$), coda ($-$), cerchio ($\circ$). Il cerchio indica incertezza tra punta e coda. L'algoritmo FCI produce un PAG eseguendo test di indipendenza condizionale e applicando regole di orientamento analoghe a quelle di Meek ma adattate alla presenza di archi bidirezionali e all'incertezza sul tipo di estremità.

Appendice C — Algebra delle equazioni strutturali: forma ridotta, identificazione e condizioni di rango

C.1 — Sistemi di equazioni strutturali simultanee

La tradizione econometrica ha sviluppato, parallelamente e in parte anticipando il programma pearliano, una teoria rigorosa dei *sistemi di equazioni strutturali simultanee* (*Simultaneous Equations Systems*, SES), la cui architettura fornisce il fondamento algebrico per molti risultati di identificazione causale. Il punto di avvio è il modello lineare strutturale in forma matriciale.

Sia $\mathbf{V} = (V_1, \ldots, V_n)^T$ il vettore delle variabili endogene, $\mathbf{X} = (X_1, \ldots, X_k)^T$ il vettore delle variabili esogene o predeterminate (equivalenti alle variabili esogene $\mathbf{U}$ nel caso deterministico, o a regressori osservati nei sistemi ibridi), e $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ il vettore dei termini di errore strutturale. Il sistema di equazioni strutturali in *forma strutturale* è:

$$\mathbf{B} \cdot \mathbf{V} + \mathbf{\Gamma} \cdot \mathbf{X} = \boldsymbol{\varepsilon}$$

dove $\mathbf{B}$ è una matrice $n \times n$ dei coefficienti strutturali tra variabili endogene (con la convenzione che la diagonale di $\mathbf{B}$ è la matrice identità — ogni variabile ha coefficiente 1 su se stessa nella propria equazione — oppure, in un'altra convenzione, la diagonale di $\mathbf{B}$ ha zeri e le equazioni hanno espliciti i termini $B_{ij} \cdot V_j$ per $i \neq j$), e $\mathbf{\Gamma}$ è una matrice $n \times k$ dei coefficienti strutturali delle variabili esogene su quelle endogene.

Nel modello ricorsivo (DAG aciclico), $\mathbf{B}$ è triangolare (inferiore o superiore, a seconda della convenzione di indicizzazione), e il sistema ha sempre soluzione unica per ogni valore di $\mathbf{X}$ e $\boldsymbol{\varepsilon}$:

$$\mathbf{V} = -\mathbf{B}^{-1}\,\mathbf{\Gamma}\,\mathbf{X} + \mathbf{B}^{-1}\,\boldsymbol{\varepsilon} = \mathbf{\Pi}\,\mathbf{X} + \mathbf{v}$$

dove $\mathbf{\Pi} = -\mathbf{B}^{-1}\,\mathbf{\Gamma}$ è la *matrice dei coefficienti ridotti* e $\mathbf{v} = \mathbf{B}^{-1}\,\boldsymbol{\varepsilon}$ è il vettore degli errori in forma ridotta. Questa è la *forma ridotta* del sistema strutturale: esprime ogni variabile endogena come funzione lineare delle sole variabili esogene.

C.2 — Il problema dell'identificazione dei parametri strutturali

Il problema fondamentale dell'algebra dei sistemi strutturali simultanei è l'*identificazione*: è possibile recuperare i parametri strutturali originali $(\mathbf{B}, \mathbf{\Gamma})$ a partire dalla forma ridotta $\mathbf{\Pi}$? In altri termini, il sistema $\mathbf{\Pi} = -\mathbf{B}^{-1}\,\mathbf{\Gamma}$ ha soluzione unica per $\mathbf{B}$ e $\mathbf{\Gamma}$ dato $\mathbf{\Pi}$?

In generale, la risposta è negativa: molteplici coppie $(\mathbf{B}, \mathbf{\Gamma})$ possono produrre la stessa forma ridotta $\mathbf{\Pi}$, perché la trasformazione $\mathbf{\Pi} = -\mathbf{B}^{-1}\,\mathbf{\Gamma}$ è una funzione molti-a-uno. Ogni *matrice di trasformazione* non singolare $\mathbf{A}$ di dimensioni $n \times n$ produce un sistema alternativo $(\mathbf{AB}, \mathbf{A\Gamma})$ con la stessa forma ridotta:

$$-(\mathbf{AB})^{-1}\,(\mathbf{A\Gamma}) = -\mathbf{B}^{-1}\,\mathbf{A}^{-1}\,\mathbf{A}\,\mathbf{\Gamma} = -\mathbf{B}^{-1}\,\mathbf{\Gamma} = \mathbf{\Pi}$$

L'insieme di tutte le matrici $\mathbf{A}$ che lasciano invariata la forma ridotta — chiamato *gruppo di osservabilità* o *insieme delle trasformazioni ammissibili* — è la sorgente della non-identificazione nei sistemi simultanei.

L'identificazione di una singola equazione strutturale — diciamo la i-esima equazione, corrispondente alla i-esima riga di $\mathbf{B}$ e $\mathbf{\Gamma}$ — richiede l'imposizione di *restrizioni* sui parametri: restrizioni di esclusione (alcune variabili esogene o endogene non compaiono in quella equazione, ovvero alcuni coefficienti sono vincolati a zero), restrizioni sui segni o sui valori, o restrizioni di covarianza sugli errori.

Le condizioni di identificazione classiche per i sistemi strutturali simultanei sono la *condizione d'ordine* e la *condizione di rango*.

C.3 — Condizione d'ordine e condizione di rango

Consideriamo l'i-esima equazione strutturale. Sia:

n = numero totale di variabili endogene; k = numero totale di variabili esogene; n_i = numero di variabili endogene che compaiono nell'equazione i (con coefficiente non zero); k_i = numero di variabili esogene che compaiono nell'equazione i.

Le variabili escluse dall'equazione i sono: $(n - n_i)$ variabili endogene escluse e $(k - k_i)$ variabili esogene escluse.

Condizione d'ordine (condizione necessaria per l'identificazione):

$$(k - k_i) \geq (n_i - 1)$$

Il numero di variabili esogene escluse dall'equazione deve essere almeno uguale al numero di variabili endogene incluse nell'equazione meno una. Intuitivamente, ogni variabile endogena aggiuntiva nell'equazione richiede uno "strumento" (una variabile esogena esclusa) per identificare il suo coefficiente strutturale.

Se $(k - k_i) = (n_i - 1)$, l'equazione è *esattamente identificata*: esiste un'unica soluzione per i parametri strutturali. Se $(k - k_i) > (n_i - 1)$, l'equazione è *sovra-identificata*: esistono più vie per stimare i parametri strutturali, e la coerenza tra di esse può essere testata (test di Sargan, test J di Hansen). Se $(k - k_i) < (n_i - 1)$, l'equazione è *sotto-identificata* e i parametri non sono recuperabili.

La condizione d'ordine è necessaria ma non sufficiente: è possibile che la condizione numerica sia soddisfatta ma che le restrizioni non siano "informative" nel senso algebrico. La condizione sufficiente (e necessaria) è la condizione di rango.

Condizione di rango (condizione necessaria e sufficiente per l'identificazione):

Sia $\mathbf{M}_i$ la sottomatrice di $\mathbf{\Pi}$ (la matrice dei coefficienti ridotti) ottenuta selezionando le righe corrispondenti alle variabili endogene escluse dall'equazione i e le colonne corrispondenti alle variabili esogene escluse dall'equazione i. La condizione di rango richiede:

$$\text{rk}(\mathbf{M}_i) = n_i - 1$$

ovvero la sottomatrice deve avere rango pieno (uguale al numero di variabili endogene incluse meno una). Questa condizione garantisce che le restrizioni di esclusione siano linearmente indipendenti — che ogni variabile esogena esclusa "strumenta" una diversa combinazione lineare delle variabili endogene incluse.

La dimostrazione della necessità e sufficienza della condizione di rango procede come segue. La soluzione per il vettore dei coefficienti strutturali della i-esima equazione, chiamato **b**_i (un vettore di lunghezza n_i), è dato dal sistema:

M_i · **b**_i = **0**

(le restrizioni di esclusione delle variabili esogene impongono che la proiezione dei coefficienti ridotti sulle variabili escluse sia zero per la i-esima equazione). Questo sistema ha soluzione unica non banale — ovvero **b**_i è identificato a meno di una costante di normalizzazione — se e soltanto se il rango di **M**_i è esattamente n_i − 1, che è esattamente la condizione di rango. □

C.4 — Variabili strumentali e stima dei parametri strutturali

La condizione di rango produce non solo un criterio di identificabilità ma anche una procedura di stima: il metodo delle *variabili strumentali* (IV). Nell'equazione strutturale esattamente identificata per V_1, si ipotizza che V_1 = β_2 V_2 + … + β_{n_i} V_{n_i} + γ_1 X_1 + … + γ_{k_i} X_{k_i} + ε_1, dove ε_1 è correlato con V_2, …, V_{n_i} (endogeneità). Siano Z_1, …, Z_{n_i − 1} le variabili esogene escluse che fungono da strumenti: soddisfano le condizioni Cov(Z_j, ε_1) = 0 (validità) e rk(Cov(**Z**, **V**_endog)) = n_i − 1 (rilevanza, equivalente alla condizione di rango).

Lo stimatore IV in forma matriciale è:

$$\hat{\beta}_{IV} = (\mathbf{Z}^\top \mathbf{W})^{-1} \mathbf{Z}^\top \mathbf{y}$$

dove **y** è il vettore delle osservazioni di V_1, **W** è la matrice delle variabili endogene sul lato destro dell'equazione, e **Z** è la matrice degli strumenti. Questo stimatore è consistente sotto le condizioni di validità e rilevanza: $\hat{\beta}_{IV} \xrightarrow{p} \beta$ al crescere della dimensione campionaria.

Nel caso sovra-identificato — più strumenti del necessario — lo stimatore ottimale è il *Two-Stage Least Squares* (2SLS), che proietta le variabili endogene sugli strumenti nella prima fase e usa le proiezioni come regressori nella seconda fase:

Fase 1: per ogni V_j endogeno, stimare $\hat{V}_j = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top V_j$ (proiezione di V_j sullo spazio degli strumenti);

Fase 2: stimare V_1 = β_2 $\hat{V}$_2 + … + β_{n_i} $\hat{V}$_{n_i} + γ_1 X_1 + … + γ_{k_i} X_{k_i} + η per OLS.

Lo stimatore 2SLS è equivalente al GMM (*Generalized Method of Moments*) con matrice di ponderazione ottimale nel caso omoschedastico, ed è asintoticamente normale con varianza stimabile dagli errori standard standard della seconda fase (con correzione per la prima fase).

C.5 — Connessione con il do-calculus nel caso lineare

Nel caso lineare ricorsivo, la connessione tra la teoria delle equazioni strutturali e il do-calculus è particolarmente trasparente. L'effetto causale totale di V_i su V_j — ovvero $\partial \mathrm{E}[V_j \mid \mathbf{do}(V_i = v)] / \partial v$ — è dato dalla somma degli effetti lungo tutti i cammini causali orientati da V_i a V_j nel DAG, ovvero dalla *regola del cammino* di Wright:

$$\partial \mathrm{E}[V_j \mid \mathbf{do}(V_i = v)] / \partial v = \Sigma_{\text{cammini } p: V_i \rightsquigarrow V_j} \prod_{(V_a \to V_b) \in p} \beta_{ba}$$

Questo risultato è derivabile direttamente dalla forma ridotta: l'effetto interventistico di V_i su V_j è l'elemento (j, i) della matrice dei coefficienti ridotti del sistema mutilato — il sistema in cui l'equazione di V_i è sostituita da V_i := v. Poiché la mutilazione dell'equazione di V_i equivale a eliminare la riga i dalla matrice **B** e a sostituirla con la riga identità e_i^T, il sistema mutilato ha matrice strutturale $\mathbf{B}^{(i)}$ con $\mathbf{B}^{(i)}_{ii} = 1$ e $B^{(i)}_{ij} = 0$ per $j \neq i$, e la forma ridotta del sistema mutilato è:

$$\mathbf{\Pi}^{(i)} = -(\mathbf{B}^{(i)})^{-1} \mathbf{\Gamma}^{(i)}$$

L'effetto interventistico è estratto dalla riga j della matrice $(\mathbf{B}^{(i)})^{-1}$, che si calcola esplicitamente usando la formula di inversione per matrici triangolari (nel caso ricorsivo) o la formula di Sherman-Morrison-Woodbury per la perturbazione di rango 1 (nel caso generale ricorsivo):

$$[(\mathbf{B}^{(i)})^{-1}]_{jk} = B^{-1}_{jk} - [\mathbf{B}^{-1}]_{ji} B^{-1}_{ik} / [\mathbf{B}^{-1}]_{ii}$$

La regola del cammino di Wright è dunque un corollario della forma ridotta del modello strutturale lineare, e l'intero do-calculus nel caso lineare è riducibile all'algebra delle matrici dei coefficienti strutturali. Questo mostra che la teoria di Pearl generalizza in modo non banale la tradizione econometrica: il do-calculus vale per sistemi non lineari, con variabili discrete, con strutture di confondimento complesse, tutti casi in cui l'algebra matriciale non è applicabile.

Appendice D — Teoria della probabilità condizionale, distribuzioni di intervento e proprietà asintotiche degli stimatori causali

D.1 — Misure di probabilità condizionale e regolarità

Si considera uno spazio di probabilità $(\Omega, \mathcal{F}, \mathrm{P})$ e un insieme finito di variabili aleatorie $\mathbf{V} = (V_1, \ldots, V_n)$ a valori in spazi misurabili $(\mathcal{V}_1, \mathcal{B}_1), \ldots, (\mathcal{V}_n, \mathcal{B}_n)$. La distribuzione congiunta di **V** è la misura di probabilità su $(\mathcal{V}_1 \times \ldots \times \mathcal{V}_n, \mathcal{B}_1 \otimes \ldots \otimes \mathcal{B}_n)$ indotta da **V**.

La *probabilità condizionale* $P(V_i \in A \mid \mathbf{V}_{-i} = \boldsymbol{v}_{-i})$ — la probabilità di $V_i \in A$ dato il valore delle restanti variabili — è definita come la versione regolare della probabilità condizionale rispetto alla σ-algebra generata da $\mathbf{V}_{-i}$. L'esistenza e unicità (P-q.c.) di tale versione regolare è garantita per spazi polacchi (spazi metrici separabili e completi) dalla teoria delle probabilità condizionali di Doob-Kolmogorov.

La distribuzione interventistica $P(\cdot \mid \mathbf{do}(\mathbf{x}))$ è, dal punto di vista misura-teorico, la distribuzione indotta dal modello mutilato $\mathbf{M}_{\mathbf{x}}$: è una misura di probabilità sul medesimo spazio misurabile delle variabili osservate, ma distinta da P perché generata da un sistema di equazioni strutturali modificato. Crucialmente, $P(\cdot \mid \mathbf{do}(\mathbf{x}))$ non è una probabilità condizionale nel senso misura-teorico standard — non è la distribuzione di $\mathbf{V}$ condizionata all'evento $\{\mathbf{X} = \mathbf{x}\}$ sotto P — ma una distribuzione definita su un modello diverso. Questa distinzione è il cuore della separazione tra associazione e intervento nella scala causale di Pearl.

D.2 — Consistenza e normalità asintotica dello stimatore di aggiustamento

Lo stimatore dell'effetto causale medio basato sulla formula di aggiustamento per il backdoor è:

$$\mathrm{ATE}_{\mathrm{adj}} = (1/N) \sum_{i=1}^{N} [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$$

dove $\hat{\mu}_t(x) = \hat{E}[Y \mid T = t, X = x]$ è un stimatore non parametrico (o semi-parametrico) delle funzioni di regressione condizionale. Sotto le ipotesi di: (a) ignorabilità forte, (b) overlap — $\exists\ \varepsilon > 0$ tale che $\varepsilon < P(T = 1 \mid X) < 1 - \varepsilon$ P-q.c. — e (c) consistenza degli stimatori $\hat{\mu}_t$ per le funzioni di regressione vere μ_t, lo stimatore $\mathrm{ATE}_{\mathrm{adj}}$ è consistente per ATE.

La normalità asintotica di $\sqrt{N}\,(\mathrm{ATE}_{\mathrm{adj}} - \mathrm{ATE})$ con varianza σ^2_{adj} si ottiene applicando il teorema centrale del limite funzionale (CLT di Donsker) al processo empirico associato a $(\hat{\mu}_1(X) - \hat{\mu}_0(X))$, richiedendo che le classi funzionali $\{\mu_t(\cdot) : t \in \{0,1\}\}$ soddisfino le condizioni di Donsker (o che si usino stimatori parametrici con convergenza $\sqrt{N}$-parametrica).

La varianza asintotica σ^2_{adj} ha la forma:

$$\sigma^2_{\mathrm{adj}} = \mathrm{Var}[\mu_1(X) - \mu_0(X)] + E[\mathrm{Var}(Y \mid T=1, X) / P(T=1 \mid X)] + E[\mathrm{Var}(Y \mid T=0, X) / (1 - P(T=1 \mid X))]$$

Il primo termine è la varianza dell'effetto eterogeno (variabilità dell'effetto causale individuale attraverso le covariate), il secondo e il terzo sono termini di errore di stima nei due gruppi di trattamento. Questa decomposizione mostra che la varianza asintotica dell'estimatore di aggiustamento è minimizzata quando il propensity score è vicino a 0.5 per tutti gli individui (bilanciamento ottimale) e massimizzata nei casi di sovrapposizione scarsa.

D.3 — La proprietà di doppia robustezza

Lo stimatore *doppiamente robusto* (*Doubly Robust*, DR) combina il modello di outcome e il modello di propensity score in modo tale che sia consistente se almeno uno dei due modelli è correttamente specificato:

ATE_DR = (1/N) Σ_{i=1}^N [μ̂_1(X_i) − μ̂_0(X_i) + T_i(Y_i − μ̂_1(X_i)) / ê(X_i) − (1−T_i)(Y_i − μ̂_0(X_i)) / (1 − ê(X_i))]

dove ê(X) = P̂(T = 1 | X) è lo stimatore del propensity score.

La doppia robustezza segue dal fatto che il bias dello stimatore DR è:

E[ATE_DR] − ATE = E[(μ_1(X) − μ̂_1(X)) · (1 − ê(X)/e(X))] + E[(μ̂_0(X) − μ_0(X)) · ê(X)/(1−e(X)) · (1 − e(X)/(1−ê(X)))]

dove e(X) = P(T=1|X) è il propensity score vero. Se il modello di outcome è correttamente specificato, μ̂_t = μ_t e entrambi i termini si annullano. Se il modello di propensity score è correttamente specificato, ê(X) = e(X) e i fattori moltiplicativi (1 − ê/e) e (1 − e/(1−ê)) si annullano. In entrambi i casi il bias è zero, dunque lo stimatore è consistente. □

La proprietà di doppia robustezza è connessa alla teoria del *Targeted Learning* di van der Laan e Rubin (2006), che costruisce stimatori efficienti nel senso della teoria semiparametrica (raggiungono il limite di Cramér-Rao semiparametrico) combinando stimatori del propensity score e del modello di outcome con aggiornamenti *targeting* mirati a ridurre il bias del funzionale causale di interesse.

D.4 — Bound di Balke-Pearl: derivazione via programmazione lineare

Si riprende qui, in forma più dettagliata, la derivazione dei limiti causali per variabili binarie in presenza di confondente latente non osservato. Siano X, Y ∈ {0, 1} variabili binarie, con X trattamento e Y esito, e sia U ∈ {0, 1} un confondente latente binario che influenza sia X che Y. Il DAG ha archi U → X, U → Y, X → Y.

Per la natura binaria di tutte le variabili, le probabilità osservabili sono:

p_{yx} = P(Y = y, X = x) per y, x ∈ {0, 1}

con la vincolo p_{00} + p_{01} + p_{10} + p_{11} = 1. Le probabilità latenti sono:

π_{xu} = P(X = x | U = u) e q_u = P(U = u) per x, u ∈ {0, 1}

e le probabilità condizionali dell'esito:

r_{y|xu} = P(Y = y | X = x, U = u)

Le quantità di interesse interventistiche sono:

$P(Y = 1 \mid \mathbf{do}(X = x)) = \Sigma_u P(Y = 1 \mid X = x, U = u) \cdot P(U = u) = \Sigma_u r_{1|xu} \cdot q_u$

L'ACE è dunque:

$ACE = \Sigma_u (r_{1|1u} - r_{1|0u}) \cdot q_u$

Per derivare i bound, si osserva che le probabilità osservabili p_{yx} sono funzioni delle probabilità latenti (π_{xu}, $r_{y|xu}$, q_u) attraverso la marginalizzazione su U:

$p_{yx} = \Sigma_u r_{y|xu} \cdot \pi_{xu} \cdot q_u$

Questo sistema — nel caso binario — produce quattro equazioni (per $(y,x) \in \{0,1\}^2$) in un numero maggiore di incognite. I bound sull'ACE si ottengono massimizzando e minimizzando ACE soggetto alle quattro equazioni di compatibilità e ai vincoli di non-negatività e normalizzazione su tutte le probabilità latenti. Questo è un *programma lineare* (l'ACE è lineare nelle incognite r e q, e i vincoli sono lineari) la cui soluzione ottimale fornisce i bound più stretti (*sharp bounds*):

$ACE_{upper} = \min\{P(Y=1, X=1) + P(Y=0, X=0), P(Y=1)\} - \max\{0, P(Y=1) - P(X=1), P(X=0) - P(Y=0)\}$

$ACE_{lower} = \max\{P(Y=1, X=1) + P(Y=0, X=0) - 1, P(Y=1) - P(X=1), P(X=0) - P(Y=0)\} - \min\{P(Y=1, X=1), P(Y=0, X=0)\}$

(Questi sono i *natural bounds* di Balke e Pearl per il caso senza strumenti aggiuntivi.)

Quando è disponibile uno strumento Z — una variabile che influenza X senza influenzare direttamente Y — le equazioni di compatibilità si arricchiscono (una per ogni valore di Z) e i bound si restringono corrispondentemente. I *bounds di Balke-Pearl con strumento* possono essere notevolmente più stretti dei natural bounds, e in alcuni casi collassano sull'ACE puntuale (identificazione completa), recuperando il risultato dello stimatore IV nel caso non-parametrico.

D.5 — Consistenza del test di d-separazione e analisi della potenza

L'analisi statistica della d-separazione — il test della domanda "$V_i \perp\!\!\!\perp V_j \mid \mathbf{Z}$?" a partire da dati campionari finiti — è un problema di test di indipendenza condizionale. Nella pratica degli algoritmi di scoperta causale (PC, FCI), si usano tipicamente:

Test chi-quadrato (per variabili discrete): la statistica $X^2 = N \cdot \Sigma_{i,j,\mathbf{z}} (p_{ij|\mathbf{z}} - p_{i|\mathbf{z}}\, p_{j|\mathbf{z}})^2 / (p_{i|\mathbf{z}}\, p_{j|\mathbf{z}}\, p(\mathbf{z}))$ converge sotto l'ipotesi nulla di indipendenza a una distribuzione chi-quadrato con $(|\mathrm{dom}(V_i)|-1)(|\mathrm{dom}(V_j)|-1)$ gradi di libertà per ogni condizionamento $\mathbf{z}$, con correzione per il numero di livelli dell'insieme di condizionamento $\mathbf{Z}$.

Test di correlazione parziale (per variabili continue gaussiane): sotto la distribuzione normale multivariata, $V_i \perp\!\!\!\perp V_j \mid \mathbf{Z}$ se e soltanto se la correlazione parziale $\rho_{ij|\mathbf{Z}} = 0$. Il test usa la statistica:

$$t = \hat{\rho}_{ij|\mathbf{Z}} \cdot \sqrt{(N - |\mathbf{Z}| - 2)} \,/\, \sqrt{(1 - \hat{\rho}^2_{ij|\mathbf{Z}})}$$

che è distribuita come una t di Student con $N - |\mathbf{Z}| - 2$ gradi di libertà sotto l'ipotesi nulla, dove N è la dimensione campionaria e $|\mathbf{Z}|$ è la cardinalità dell'insieme di condizionamento.

La potenza del test decresce all'aumentare di $|\mathbf{Z}|$ — condizionare su molte variabili riduce i gradi di libertà effettivi — il che rende i test di indipendenza condizionale su insiemi di condizionamento grandi molto conservativi e poco informativi. Questo fenomeno è uno dei limiti principali degli algoritmi di scoperta causale basati su test di indipendenza condizionale: per grafi densi, dove gli insiemi di separazione possono essere grandi, la potenza dei test può essere insufficiente anche per dimensioni campionarie molto elevate.

La *consistenza* dell'algoritmo PC — nel senso che, al crescere di N, l'algoritmo produce il CPDAG corretto con probabilità tendente a 1 — richiede che il livello dei test di indipendenza $\alpha = \alpha(N)$ tenda a 0 con una velocità appropriata rispetto a N. Condizioni sufficienti per la consistenza sono state derivate da Kalisch e Bühlmann (2007) nel contesto dei modelli gaussiani, mostrando che per grafi con p nodi e massimo grado causale q, l'algoritmo PC con $\alpha(N) = 1/N^{1/(4q+2)}$ è consistente in alta dimensione ($p \to \infty$ con $N \to \infty$) sotto condizioni di regolarità sulla matrice di covarianza.

Il tasso di convergenza implica che per ogni coppia di nodi non adiacenti nel grafo vero, la probabilità che l'algoritmo li colleghi erroneamente — un *falso positivo strutturale* — tende a zero al tasso $N^{-1/(4q+2)}$, mentre la probabilità di non collegare due nodi adiacenti — un *falso negativo strutturale* — richiede condizioni aggiuntive sulla forza delle correlazioni parziali (bounded away from zero uniformly), solitamente espresse nella forma $\rho_{ij|\mathbf{Z}} \geq c > 0$ per qualche costante c indipendente da N.